Native Language Identification: Explorations and Applications

Shervin Malmasi

Bachelor of Computer Science (Hons Class I), The University of Sydney
Bachelor of Arts (Linguistics, Spanish), The University of Sydney
Bachelor of Science (Psychology), The University of Sydney

This dissertation is presented for the degree of

Doctor of Philosophy

at

MACQUARIE University
SYDNEY·AUSTRALIA

CENTRE FOR LANGUAGE TECHNOLOGY

February 2016

Abstract

The prediction of an author's native language using only their second language writing — a task called Native Language Identification (NLI) — is usually tackled using supervised classification. This is underpinned by the presupposition that an author will be disposed towards certain language production patterns in their second language (L2), due to influence from their first language (L1). This identification of such L1-specific linguistic patterns can be used to study language transfer effects in Second Language Acquisition (SLA). NLI can also be used in forensic linguistics, serving as a tool for authorship profiling to provide evidence about a writer's linguistic background.

NLI is a young but rapidly growing research topic. It has become a well-defined classification task in the past decade and increasing work during the last five years has brought an unprecedented level of research focus and momentum to the area, culminating in the first NLI Shared Task in 2013. Most work hitherto has focused on the core machine learning and feature engineering facets of the task, obtaining suitable data and unifying the area with a common evaluation framework.

This thesis makes three broad contributions: (1) exploring the task in new ways; (2) investigating how NLI can inform SLA; and (3) introducing the novel task of L1-based text segmentation.

Following our implementation of an NLI system for the shared task — which investigated the effects of classifier ensembles, feature types and feature diversity — we explored the task in several new ways. We first looked at human and oracle performance, gauging potential for further improvements in classification performance. We used two oracles to estimate upper bounds for NLI accuracy, applying them to a new dataset composed of all submissions to the 2013 shared task, revealing interesting error patterns. We then presented the first study of human performance for NLI using a group of experts. The experts did not outperform our NLI system, with the performance gap likely to widen on the standard NLI setup, demonstrating that this is a hard task that, uncharacteristically, computers can perform better than humans. We next explored the cross-lingual applicability of NLI by extending it to other languages. To this end we identified six typologically very different sources of non-English L2 data and via a series of experiments using common features established that NLI accuracy is similar across the L2s and a wide range of L1s. We showed that other patterns, *e.g.* oracle performance and feature diversity, also hold across languages.

Next, we considered practical applications of NLI in SLA research, investigating ways to use the classification task to give a broad linguistic interpretation of the data. Our first exploration focused on language transfer, the characteristic L2 usage patterns caused by native language interference, which is investigated by SLA researchers seeking to find overused and underused features. We proposed a method for deriving ranked lists of such discriminative features and then analyzed our results to see how useful they might be in formulating plausible language transfer hypotheses. We then defined and examined an approach to formulating and testing hypotheses about errors and the environments in which they are made, a process which traditionally in SLA often involves substantial effort. To this end we defined a new task for finding contexts for errors that vary with the native language of the writer and propose four graph-theoretic models for doing so. The findings in this chapter form the basis of a useful research direction for developing methods to assist SLA experts develop hypotheses using large data.

The final part of this dissertation introduced the novel task of native language-based text segmentation, exploring how discriminative NLI features investigated in the previous task can be exploited here. The goal is to partition a text into regions that exhibit differing L1 influence; such methods could be applied for intrinsic plagiarism detection or even literary analysis. We adapted an unsupervised Bayesian approach originally developed for topic segmentation to one with generative models built over features useful in NLI. We investigated several models: one with alternating asymmetric priors was the best, with compactness of distributions over features proving to be important.

Contents

List of Tables

List of Figures

Declaration

The research presented in this thesis is the original work of the author except where otherwise indicated. This work has not been submitted for a degree or any other qualification to any other university or institution. All verbatim extracts have been distinguished by quotations, and all sources of information have been specifically acknowledged.

Some parts of this thesis include revised versions of published papers:

- Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. NLI Shared Task 2013: MQ Submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1716`

- Shervin Malmasi and Mark Dras. Chinese Native Language Identification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-14)*, pages 95–99, Gothenburg, Sweden, April 2014b. Association for Computational Linguistics. URL `http://aclweb.org/anthology/E14-4019`

- Shervin Malmasi and Mark Dras. Language Transfer Hypotheses with Linear SVM Weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1385–1390, Doha, Qatar, October 2014d. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D14-1144`

- Shervin Malmasi and Mark Dras. Finnish Native Language Identification. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 139–144, Melbourne, Australia, November 2014c. URL `http://www.aclweb.org/anthology/U14-1020`

- Shervin Malmasi and Mark Dras. From Visualisation to Hypothesis Construction for Second Language Acquisition. In *Proceedings of TextGraphs-9: the Workshop on Graph-based Methods for Natural Language Processing*, pages 56–64, Doha, Qatar, October 2014f. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W14-3708`

- Shervin Malmasi and Mark Dras. Arabic Native Language Identification. In *Proceedings of the Arabic Natural Language Processing Workshop (EMNLP 2014)*, pages 180–186, Doha, Qatar, October 2014a. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W14-3625`

- Shervin Malmasi and Mark Dras. A Data-driven Approach to Studying Given Names and their Gender and Ethnicity Associations. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 145–149, Melbourne, Australia, 2014e. URL `http://www.aclweb.org/anthology/U14-1021`

- Shervin Malmasi and Mark Dras. Large-scale Native Language Identification with Cross-Corpus Evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2015)*, pages 1403–1409, Denver, CO, USA, June 2015a. Association for Computational Linguistics. URL `http://aclweb.org/anthology/N15-1160`

- Shervin Malmasi and Aoife Cahill. Measuring Feature Diversity in Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 49–55, Denver, Colorado, June 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-0606`

- Shervin Malmasi, Joel Tetreault, and Mark Dras. Oracle and Human Baselines for Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–178, Denver, Colorado, June 2015c. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-0620`
- Maolin Wang, Shervin Malmasi, and Mingxuan Huang. The Jinan Chinese Learner Corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 118–123, Denver, Colorado, June 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-0614`
- Shervin Malmasi, Eshrag Refaee, and Mark Dras. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, pages 209–217, Bali, Indonesia, May 2015b
- Shervin Malmasi and Mark Dras. Automatic Language Identification for Persian and Dari texts. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, pages 59–64, Bali, Indonesia, May 2015b
- Shervin Malmasi and Mark Dras. Language Identification using Classifier Ensembles. In *Proceedings of LT4VarDial - Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, Hissar, Bulgaria, 9 2015c
- Shervin Malmasi, Mark Dras, and Irina Temnikova. Norwegian Native Language Identification. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2015)*, pages 404–412, Hissar, Bulgaria, September 2015a. URL `http://www.aclweb.org/anthology/R15-1053`
- Shervin Malmasi and Mark Dras. Multilingual Native Language Identification. *Natural Language Engineering*, FirstView:1–53, December 2015d. ISSN 1469-8110. doi: 10.1017/S1351324915000406

Shervin Malmasi

Date: 22 February 2016

Acknowledgements

First and foremost, I would like to thank my supervisor, Mark Dras, for all his help and support. He provided me with valuable advice, insight and encouragement from the very first time I met him, through to the completion of this work, and beyond. He was patient throughout the project and never hesitated to provide me with the necessary time and resources. I couldn't imagine having a better mentor than Mark and I consider myself very fortunate for having had him as my advisor. I will be eternally grateful to him for all his help over the years.

A special thanks goes to the members of my thesis examination committee, Martin Chodorow, James Curran and Ani Nenkova, for their insightful and constructive comments. I would also like to thank the organizers of the 2013 Native Language Identification Shared Task: Joel Tetreault, Aoife Cahill and Daniel Blanchard. Participating in the task was a formative experience for me which also helped jumpstart this research. The dataset released as part of that task also played an important part in this work.

I would like to express my gratitude to my friends and colleagues in Macquarie University's Department of Computing and the Centre for Language Technology: Benjamin Börschinger, Mitchell Buckley, Sepehr Damavandinejadmonfared, Oldooz Dianat, Christophe Doche, Lan Du, Mary Gardiner, Matthew Honnibal, Mark Johnson, Kallol Karmakar, Anish Kumar, François Lareau, Bernard Mans, Diego Molla-Aliod, Yasaman Motazedi, Jason Naradowsky, Dat Nguyen, Mehdi Parviz, John Pate, Abeed Sarker, Rolf Schwitter, Jette Viethen, Yan Wang, Mahmood Yousefi, and Tony Zhao. A very special thank you also goes to Jojo Wong, who helped me get started with this undertaking at the beginning.

I'm also grateful to the department administrators Melina Chan, Sylvian Chow, Camille Hoffman, Donna Hua, Jackie Walsh, Fiona Yang, Jane Yang, as well as all of the Science IT staff. They work hard to keep things running smoothly so that we can focus on our research.

Outside of our lab I would like to express my thanks to Aoife Cahill, Cyril Goutte, Hamed Hassanzadeh, Irina Temnikova, Joel Tetreault, Maolin Wang, and Marcos Zampieri for their help, advice and collaboration.

I also thank all of the researchers who provided the numerous datasets used in this work. A special acknowledgement also goes to the many anonymous reviewers who provided feedback on the numerous publications produced during this research; all of their comments and critical remarks helped shape this dissertation in some way.

Finally, I dedicate this dissertation to my family. This journey would not have been possible without their unconditional love and support.

Chapter 1

Introduction

Chapter Contents

Advances in Natural Language Processing (NLP) research have often (but not always) resulted in increased cooperation and interaction between NLP and linguistics. NLP technology has had a direct impact on linguistics, resulting in the development of advanced data processing tools for applied and corpus linguists. On the other hand, linguistic research continues to inform NLP in various ways and has benefited tasks such as part-of-speech tagging, morphological analysis, speech recognition and more.

Machine Learning (ML) and Second Language Acquisition (SLA) are two major research areas related to NLP and linguistics, respectively. A specific trend tying these together has been the application of NLP and ML to study SLA and other phenomena related to language learning. As the name implies, SLA is concerned with the processes involved in acquiring a second language, including that of language transfer. This developing connection and synergy between ML and SLA can help provide many insights into learner language, and is a primary motivation of this thesis.

It is well known that the manner in which most non-native speakers use a language is strongly influenced by their native or mother tongue. The most obvious manifestation of this comes from speech where speakers' accents are influenced by the distinct phonology of their native tongue. However, it has been shown that non-native writings also contain discriminative signals about the writer's linguistic background. Just as a non-native speaker can be identified by his or her accent, influence from an individual's mother tongue leads to the systematic manifestation of language use patterns and other linguistic phenomena that can reliably help identify the native language of the author of a particular text.

It has been observed in the linguistics literature since at least the 1950s that native speakers of particular languages make characteristic errors when speaking or writing in a second language. For example, an erroneous sentence such as *the development of country park can directly alleviate overcrowdedness* is very unlikely to occur in the English written by a Spanish native speaker, whereas it is more likely to be written by a Chinese native speaker. This is due to the fact that determiner errors such as this are more likely to be committed by Chinese speakers. This phenomenon has been investigated independently in a number of different fields from different perspectives, including qualitative and quantitative research in SLA and Foreign Language Teaching and Learning (FLTL), and more recently though predictive computational models in NLP; yet until now, there has been surprisingly little synergy between the various approaches.

In the field of SLA, an early focus was on the observable errors made by speakers of a particular native language (called the L1) in speaking or writing in another language (the L2); *e.g.*, in the Contrastive Analysis (CA) approach (Lado, 1957). Much of the motivation here was and continues to be pedagogical, with the aim of identifying good practices in FLTL. Identifying characteristic errors of different native speakers helps with providing appropriate feedback to learners and with creating appropriate teaching materials. This has become a key topic within the field, and as Ellis (2008) notes in his comprehensive overview of SLA, the issue of identifying which effects on L2 acquisition are attributable to L1 and which to general developmental (or other) processes is a pervasive one. Similarly, Ortega (2009) gives as a key question of interest "What is the role played by first language in L2 development, *vis-à-vis* the role of other universal development forces?"

Such transfer effects are generally studied and identified through the comparison of the L2 writings of learners from two or more L1 backgrounds. This methodology can identify phenomena such as avoidance and overuse; it is also able to minimise overestimation of transfer effects. The difficulty, then, is that in SLA research most identification of candidate transfer effect constructions is done manually, and it is difficult to carry this out for a large amount of text. Small quantities of text can lead to poor estimates of the true extent of underlying cross-linguistic transfer effects, just by the nature of statistical sampling. Ellis cites seven studies which give proportions of cross-linguistic errors ranging from 3% to 51%. While some of this extreme variability is attributable to differences in methodology, sample size would be a major factor. A large-scale study would therefore minimise this problem.

Granger (2011) discusses the methodology of carrying out corpus-based approaches that use some automated error identification techniques as a solution to this. They show some promise, but tend to be few in number and limited in scope. An example is the study of Diéz-Bedmar and Papp (2008), comparing Chinese and Spanish learners of English with respect to the English article system (*a*, *an*, *the*). Drawing on 175 texts, they take a particular theoretical analysis (the so-called Bickerton semantic wheel), use the simple Wordsmith tools designed to extract data for lexicographers to identify errors in a semi-automatic way, and evaluate using hypothesis testing (chi-square and z-tests, in their case). In contrast, a fully automatic techniques would mean that — in addition to being able to process more data — any change in assumptions or in theoretical approach could be made easily, without need for manual re-annotation.

The issue of identifying L1 effects in L2 texts has been treated differently within the field of NLP, where it is typically tackled as part of the task of Native Language Identification (NLI), a subtype of text classification where the goal is to determine the native language of an author writing in a second

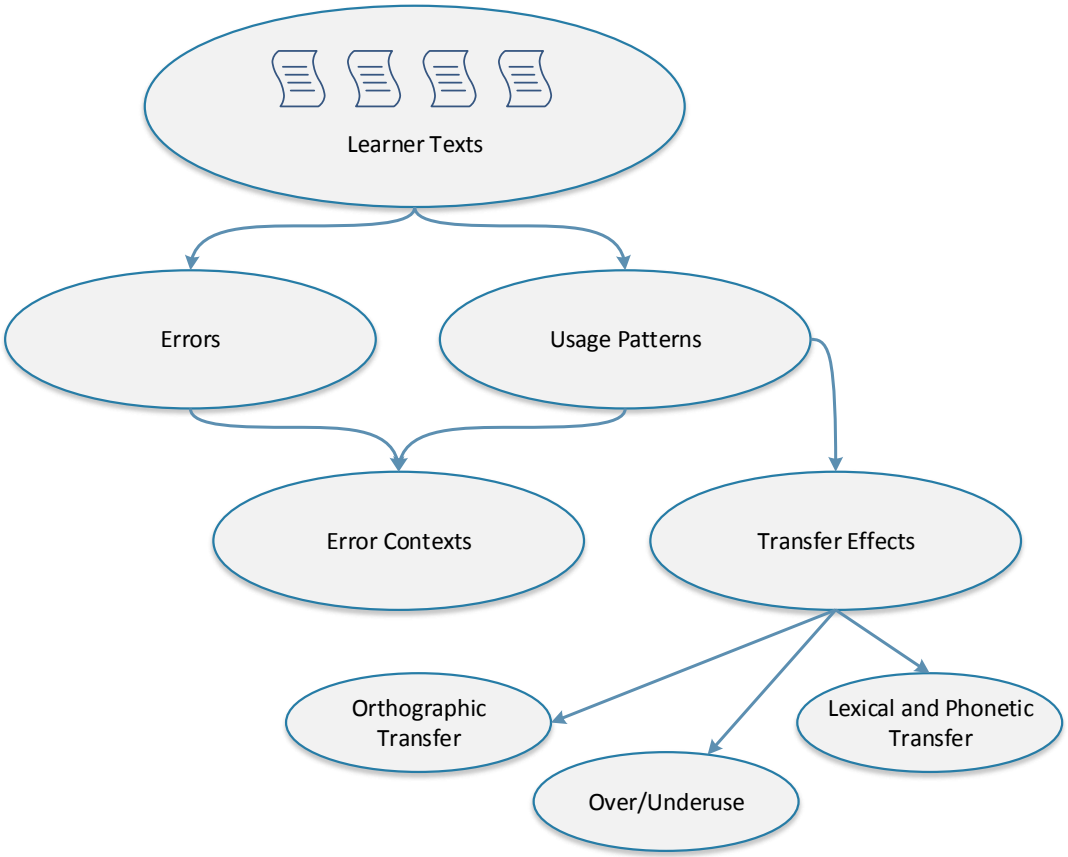

Figure 1.1: Examples of SLA-relevant information that can be extracted from learner corpora.

language. This is sometimes part of a broader task where an author is classified with respect to characteristics such as gender, age, native language, and so on. These variables are often of interest to forensic linguistics experts and intelligence agencies, as we discuss in §1.1.2. Additionally, this profile information is often of interest to marketing organisations for product promotional reasons, or to governments for crime investigation purposes, such as for phishing (Myers, 2007), the attempt to defraud through texts that are designed to deceive Internet users into giving away confidential details.

Most of the existing work on NLI treats the problem as a classification one, and adopts a supervised machine learning approach. This statistical methodology can provide insight into how individuals use their implicit and explicit knowledge about the phonology, syntax and orthography of their native language to produce text in a second language.

Recent work in this area (Wong, 2012) has demonstrated that NLI can inform us about learner language mistakes and usage patterns. Such systems could be used to automatically identify language features that are strongly associated with writers of a particular L1 background and use them in a number of ways. This includes the formulation and testing of hypotheses relating to transfer effects, but could also be expanded to more sophisticated tasks such as linking specific usage patterns and errors together to better understand the contexts in which learners make certain mistakes. Figure 1.1 depicts some of the SLA-relevant information that could be extracted from learner texts. To this end, the use of state-of-the-art tools and techniques from NLP enable us to process large quantities of data (thousands of times the size of the manual studies) in a way that will let us explore a variety of assumptions and theoretical analyses.

In a similar manner to which the development of electronic learner corpora was considered "a revolution in applied linguistics" (Granger, 1994), automated methods for extracting L1-based persistent language transfer effects via NLI are an innovative approach to tackling problems that SLA researchers have been looking at for a long time. As Milton and Tsang (1993) noted:

> Because interlanguage research is still largely pursued by manual analysis, EFL professionals lack sufficient evidence to quantify adequately students' problems in written expression. Nor do we have a reliable measure of the significance of written error. To adapt a Tang dynasty poem by Wang Wei: we work in the midst of mountains of linguistic data with no idea of their true magnitude. We have no quantitative model of the characteristics of our students' writing, and so have no real measure of the degree of specific communicative problems in varieties of learner English. Without a reliable index of the degree of difficulty that our students have with the various dimensions of written English such as its lexis, syntax, pragmatics and semantics, we are left to make do with approximations based on impressions, anecdotes and manual counts of small samples.
>
> If the corpus-linguistic techniques which have been employed so successfully to NS [Native Speaker] writing can be used, with modification, to assist in the analysis of NNS [Non-native Speaker] writing, we can demonstrate to students, teachers and textbook writers precisely how NNS written language differs from (and is similar to) native-speaker varieties. These methods might help provide an empirical measure of the effectiveness of pedagogical techniques currently employed in teaching students to understand and approximate NS writing styles. Until suitable computer programs are refined, a considerable amount of the analysis must remain manual, but human analysis can be greatly aided with current software which allows the creation of large databases that can be retrieved and searched electronically.

With the ongoing development of systems based on sophisticated linguistic features and machine learning, work in this area is moving far beyond "bean counting at the lexical level" (Milton and Tsang, 1993). Indeed, NLI can form the basis for the development of substantially more sophisticated methods for investigating learner language using newly developed learner corpora which are unprecedented in their size.

1.1 Applications

With the proliferation of text and language technology in the last decade, there are a number of fields where NLI could be applied. Here we outline several key applications that would immensely benefit from the advancement of NLI methods.

1.1.1 Language Teaching and Learning

Today, we live in a world where there more bilingual individuals than monolinguals, but multilingualism does not automatically imply having attained full mastery of multiple languages. In fact, polyglots are rarely equally well-versed in all of their languages and very few people are as proficient in their second languages as they are in their mother tongue.

As the world continues to become a highly globalized and interconnected community, the learning of foreign languages is becoming increasingly common. It is typical in many countries for this to be a part of the formal education process, beginning in either high school or university. This is particularly so in Europe, where many countries are teaching English intensively even in primary school. Alongside academia, immigration has been another driving factor behind the rise of bilingualism. Some groups command more economic or social power than others, and those individuals who wish to join these groups are often required to learn the new host's language. This requirement is often implemented through various gate-keeping mechanisms such as language tests. These have been some of the most common reasons as to why people would experience the opportunity or need to acquire an additional language.

Since the end of World War II, the world has witnessed the ascendancy of English as its *lingua franca*. One consequence of this has been that non-English speakers from across the globe are aware of the benefits of learning English, and highly motivated to do so. While some individuals may learn languages out of their own interest, most people expect to improve their lives in some meaningful way, and this increases their motivation to do so. In fact, English is already the most commonly used language in the areas of science and business. However, the majority of English speakers are not native speakers and just like any other foreign language, there can be many challenges in mastering English.

From a socio-linguistic point of view, it has been shown that listeners will often evaluate speakers on a range of attributes based on their language use. This is often measured using language attitude tests, such as the matched-guise test, that aim to uncover subconscious attitudes listeners hold about people based on the linguistic variety that they use (Lambert, 1967). These linguistic attitudes have been known to influence workplace environments, business interactions and other social systems and processes. In fact, research shows that "non-natives tend to be downgraded in contexts ranging from the classroom to the workplace" (Eisenstein, 1983). This is generally facilitated by the fact that a speaker's language use patterns have been shown to serve as markers for evaluating their personal and socio-economic characteristics (Ball, 1983). In multi-cultural settings, this has been shown to enable listeners to elicit certain stereotypes about the speaker and potentially lead to less favourable evaluations. More examples of how such perceptions arise can be found in Giles et al. (1981) and Bilaniuk (2003).

Many learners are aware of these perceptions to some degree and seek to ameliorate these effects through adapting their language production. These modification have been captured by models such as the Communication Accommodation Theory (Giles and Ogay, 2007) which postulates that speakers tend to accommodate their language production to interlocuters that they like by converging their styles to minimize differences. Alternatively, divergence may result in the accentuation of linguistic differences in order for the interlocuters to assert their distinctiveness. In this sense, speakers are motivated by what they want to accomplish and accordingly keep their audience in mind when they make linguistic choices in order to reduce linguistic dissimilarities (Giles et al., 1991).

Furthermore, while technological advancements such as the introduction of the telephone, fax, Internet and personal computing have diminished the obstacles posed by physical distances, language proficiency remains an impediment to both business and interpersonal communication. This has lead to an increasing necessity for language learning resources, which has in turn fuelled much of the language acquisition research of the past decade.

All of this provides intrinsic motivation for many of the learners to continue improving their language skills beyond that of basic communication or working proficiency towards near-native levels. In itself, this is not an easy task, but a good starting point is to reduce those idiosyncratic language use patterns that are being caused by the influence of the native language.

Within SLA research, NLI can be used to identify the most challenging aspects of a language for learners from specific L1 backgrounds. As opposed to much of the work in SLA that takes a deductive, corpus-based approach, using NLI does not require that we start with a set of pre-existing hypotheses. Instead, we canvas the available data to identify regularities and patterns that are distinct to writers from a particular L1 background. Once the significance of these patterns have been established through statistical testing, we can then hypothesize about the underlying linguistic causes of these intergroup differences. With respect to the field, Jarvis (2012) suggests that NLI is an innovative approach to tackling problems that SLA researchers have been looking at for a long time.

Based on such language-specific models, learners can be provided with customized and specific feedback, determined by their native language. For example, algorithms based on these models could provide students with much more specific and focused feedback when used in automated writing evaluation systems (Rozovskaya and Roth, 2011).

Furthermore, researchers are interested in the nature and degree to which a native language affects the acquisition and production of other consequently learnt languages. NLI-based analyses could be used to help researchers in linguistics and cognitive science to better understand the process of second language acquisition and language transfer effects. This information can better inform researchers about how language is processed in the brain, in way that are not possible by just studying monolinguals, thereby providing us with important insights that increase our knowledge and understanding of the human language faculty.

1.1.2 Forensic Linguistics and Legal applications

The field of forensic linguistics (Gibbons, 2003; Coulthard and Johnson, 2007) — a branch of applied linguistics — involves the application of linguistic and scientific knowledge within the context of the law. It is a juncture where the legal system and linguistic stylistics intersect (Gibbons and Prakasam, 2004; McMenamin, 2002).

Research in this young field has been growing since its inception in the 1960s. A series of seminars and the establishment of the International Association of Forensic Linguists[1] launched the field into the academic conference circuit and it has continued to expand since. The Centre for Forensic Linguistics at Aston University, the first research centre dedicated to this area, was opened in 2008. These developments, coupled with the growing use of linguistic evidence, have brought the discipline into the mainstream. A more detailed outline of the field's history can be found in Blackwell (2013).

Although the field has expanded to many other strands, the application of textual analysis to disputed documents remains one of its core foci. The use of software and machine-readable corpora are also becoming increasingly common in this area.

From a forensic point of view, NLI can be a useful tool for use by intelligence and law enforcement agencies. In fact, recent NLI research such as that related to the work conducted by Perkins (2014) at The Centre for Forensic Linguistics has already attracted interest and funding from intelligence

[1] http://www.iafl.org/

agencies (Perkins, 2014, p. 17). In this context NLI can be used as a tool for Authorship Profiling (Grant, 2007) in order to provide evidence about the linguistic background of an author.

There are a number of situations where a text, such as an anonymous letter, is the central piece of evidence in an investigation. The ability to extract additional information from an anonymous text can enable the authorities and intelligence agencies to learn more about threats and those responsible for them. Clues about the native language of a writer can help investigators in determining the source of anonymous text and the importance of this analysis is often bolstered by the fact that in such scenarios, the only data available to users and investigators is the text itself. One recently studied example is the analysis of extremist related activity on the web (Abbasi and Chen, 2005).

Although linguistic profiling methods are not considered reliable enough to meet the stringent legal criteria for evidence (Perkins, 2014, p. 16), their use in legal procedures is not without precedence.

McMenamin (2002, pp. 207-233) lists a number of adjudicated cases where forensic stylistics evidence played some role in the case outcome. Some examples cases include:

- A bank received various anonymous letters advising against financing a large business project. The company provided a list of likely suspect writers.

- A large hospital received a number of anonymous letters threatening an administrator. The hospital identified a number of present and former employees as possible writers.

Another case highlighted by Blackwell (2013) is that of the 2001 anthrax attacks, where a number of threatening letters containing anthrax spores were mailed to news media outlets and public figures in the US. The emerging consensus among forensic linguists who analyzed the letters was that they were likely written by a native speaker, despite several spelling mistakes and non-standard syntax. This analysis seems to have been accurate as the FBI prepared to prosecute Bruce Edwards Ivins in 2008. Although he committed suicide before the case could proceed, he was later declared the sole culprit for the attacks.

Such forensic analyses have also been extended to SMS text messages sent from mobile phones (Grant, 2010). A notable case is the conviction of David Hodgson, where the jury was persuaded by linguistic evidence of his authorship of text messages sent from his former lover's phone after her death (Grant, 2012, p. 475). Much of the evidence used here was based on discriminating the messaging styles of the two individuals using the lexical choices and preferences.

Linguistic stylistics also played a role in the case of Ted Kaczynski, also known as the "Unabomber", who waged a nationwide bombing campaign across the US from 1978 to 1995. In 1995, the Unabomber demanded that his 35,000 word manifesto be published by the media in order for him to end his bombing campaign, a request which was granted in September of that year. The perpetrator's brother, David Kaczynski, found linguistic similarities between the manifesto and his brother's writing. He examined the idiosyncrasies by comparing the manifesto to old handwritten essays and letters written by Ted, helping to identify the perpetrator.

Error analysis has also played a role in cases involving non-native writing. Hubbard (1994) reports on the use of linguistic analysis in an extortion case where an L1 Polish speaker was charged with having written a number of extortion letters to a nationwide chain store in South Africa. During the trial the defence solicited the services of an expert witness who asked the accused to write seven compositions on various topics under test conditions.

These writings were then compared against the extortion letters through an error analysis. The defence argued that it was improbable the the accused had authored the extortion letters. The prosecution then sought the assistance of a linguist to critique this evidence, using stylometric and error analyses. This process — particularly quantitative analysis of article and spelling errors — involved comparing the defendant's writing against the extortion letters and additional compositions by other L1 Polish authors who were included. The results showed that the accused's writings were the closest match to the extortionist out of the eight candidates. This evidence suggested that the defence had overstated their argument and ultimately helped secure a conviction.

An important point to consider here is the rigorous objectivity required in applying such methods during the legal process. Here, the defence and prosecution differed in their manually-applied methodology, leading to the formation of opposing arguments. The use of computational models such as NLI for automating such analyses could help increase the objectivity of the process, provided that appropriate training data is used. At the very least, it could lead to more reproducible results. The cases described in this section rely on the same types of linguistic differences that are used in NLI and demonstrate how such systems could be used as an investigative tool.

In addition to their usage in post-hoc forensic analyses, they could also be used in conjunction with real-time monitoring and detection systems. Furthermore, the derived language profiles could also be used in some security applications where the scope of information retrieval or search space could be limited to those documents matching certain native language models or linguistic criteria.

1.1.3 Other Applications

The development of highly specific L1–L2 language transfer models could be potentially useful in many other fields of language technology, including speech recognition, machine translation, parsing and part-of-speech tagging.

Such models can also be applied to the task of automatic grammatical error detection and correction. It has already been shown that contextual models based on the user's native language can assist error correction systems (Gamon et al., 2008; Lee, 2009). Such systems include those designed for correcting grammar and spelling mistakes where the availability of more specific language profiles could help improve the effectiveness and accuracy of these systems.

Most of the above-mentioned NLP systems are trained on corpora of native English text, usually resulting in much lower performance when used with non-native data. As their use becomes more widespread, language specific models would benefit these tools and enable them to be more robust and be used with a wider range of input data.

1.2 Goals and Objectives

NLI is a young but rapidly growing area of research. It has become a well-defined classification task in the past decade and increasing work during the last five years has brought an unprecedented level of research focus and momentum to the area, culminating in the first NLI Shared Task in 2013. Most work to date has focused on the core machine learning and feature engineering facets of the task, obtaining suitable data and unifying the area with a common evaluation framework.

Bearing this in mind, this thesis focuses on three broad goals within this area. These three aims, along with more specific objectives, are listed below.

1. Exploring the NLI task in new ways to gain further insights about the task and how they affect practical applications of the methods.

   (a) Assess human performance for NLI, something which has not been done to date. We also aim to compare the performance of experts against a machine learning model to see whether this is the kind of task that is hard for computers, relative to humans or *vice versa*.

   (b) Evaluate oracle performance for estimating the upper-bounds for NLI accuracy, gauging potential for further improvements in classification performance. Results are also analyzed for interesting error patterns.

   (c) Determine how well NLI systems might perform on other data, including those from other domains and genres. We use a new dataset to conduct large-scale cross-corpus evaluation to assess this using standard features and an oracle.

   (d) Explore the cross-lingual applicability of NLI by extending it to other languages to determine if results hold across languages.

   (e) Examine suitability of the methods to discriminating between native and non-native writing in a binary classification task.

2. Investigating how NLI can inform SLA by drawing together work from both areas.

   (a) Study the possibility of using the classification task to give a broad linguistic interpretation of the data. The NLI literature has not qualitatively assessed the information being captured by NLI models and how these may be connected with SLA phenomena. If the features do correspond to SLA-related issues, they can potentially be used in SLA research.

   (b) Propose a method for deriving interpretable, ranked lists of underused and overused linguistic features from large-scale data. Such a method should ideally work with arbitrary features and provide lists for individual L1s.

   (c) Analyze results to ascertain if they can be used to develop plausible language transfer hypotheses supported by current linguistic evidence. This enables us to determine how the knowledge obtained from NLI models can inform SLA research.

   (d) Define a new task for finding linguistic contexts for errors that vary with the native language of the writer, and examine a variety of models to see which best tackles the new task.

3. Applying NLI models to other NLP tasks.

   (a) Introduce the novel task of L1-based text segmentation, where the goal is to segment texts based on L1 influence. This is useful for tracing L1 influence in texts that contain such effects from more than one language and can be applied in areas such as plagiarism detection and literary analysis.

   (b) Explore a range of Bayesian models for this, including a novel one using alternating asymmetric priors, based on adapting an unsupervised topic segmentation approach to use syntactic POS n-gram features with compact distributions.

   (c) Examine how the discriminative NLI features from classifier models can be exploited in this unsupervised task.

An overarching aim here — most directly reflected by the second and third goals — is to draw together research from the computational linguistics and language learning fields to see how they can inform each other. A key aim here is to develop models using machine learning for addressing in a unified way problems regarding second language learning in the fields of SLA/FLTL and NLP: the SLA and FLTL research can benefit from use of the tools of NLP on large-scale data, while NLP tasks can benefit from the derived models and language acquisition insights from SLA.

1.3 Motivation and Significance

There are two major ways in which the proposed work is **significant**. One is related to connecting SLA to work in NLP. As discussed above, an important strand of work in SLA is motivated by the aim of improving foreign language teaching and learning; and NLP has the tools to use large amounts of data to do this (semi-)automatically. There is a reasonable body of work within NLP that also has the goal of helping learners of a language: the work by Tetreault et al. (2010) is an example of identifying specific problems for a learner, and there is some work towards building more general models (Tetreault and Chodorow, 2009). However, these are concerned with disparate phenomena, in a sense looking at learner problems on a case-by-case basis, whereas the aim of this work is to look at the more fundamental transfer-related questions identified in SLA. Moreover, while there are some attempts in SLA to use computational approaches on larger-scale data (such as in Jarvis and Crossley (2012)), these still use fairly elementary techniques and have several shortcomings, including in approaches to corpus annotation and in the computational artefacts derived from these. Overall, there is much scope for contributing better techniques to fields dedicated to understanding and improving human learning of languages; work in NLI has been moving towards this direction. A significant contribution of this thesis is that we further explore the task of NLI in new ways, including human performance, oracles and extensions to several other languages. Building on this, we also consider how these models could be used to support and inform specific tasks in SLA.

The second way in which this thesis is significant is the connection between NLI and other NLP tasks: it will highlight the ways in which knowledge about learner writing can be used in other NLP areas. In a special journal issue on the interaction of linguistics and NLP, Ken Church, one of the founders of statistical approaches in NLP, lamented the broad disappearance of linguistics from work in NLP (Church, 2011). This thesis aims to renew the relationship between linguistics and computation, at least in this particular domain, with an expectation of more such insights. We anticipate that both of these fundamental characteristics of the proposed work will also lead to the practical benefit of a deeper understanding of cross-linguistic effects and of improved performance in NLI-related tasks.

The work is **innovative** in a number of ways. The combination of SLA and the abovementioned state-of-the-art NLP techniques is new, as are the ideas behind the proposed approaches, detailed below in the individual tasks into which the work will be broken. In terms of NLI, the experiments discussed in Chapter 5 are the first to explore the application of these methods to languages beyond English. This work is also the first to examine other issues such as human performance and feature interaction. In terms of error and language transfer detection, the application to SLA and to identifying transfer effects is itself novel. And in terms of bringing these together, the use of the graph-theoretic models we propose for identifying error contexts is a novel approach to going beyond

the use of linguists' intuition as a method for generating candidate hypotheses for an SLA study, quite different from existing applications of graph-theoretic methods in NLP. Finally, the native language-based text segmentation task proposed and tackled in Chapter 7 is also a novel contribution of this thesis.

1.4 Thesis Outline

The remainder of this thesis is organized as follows:

- In Chapter 2 we introduce key topics and a detailed overview of related work, including an exhaustive survey of the NLI literature.
- In Chapter 3 we develop our NLI system and examine relevant machine learning issues.
- Chapter 4 investigates oracles, human performance and cross-corpus evaluation for NLI.
- Chapter 5 extends our NLI methods to a range of other languages to assess their cross-lingual applicability.
- Chapter 6 discusses how the NLI methodology can be used to extract relevant features of learner language that can assist qualitative research in language teaching and acquisition.
- Chapter 7 uses our NLI work to propose an unsupervised method for detecting influence from multiple L1s within a single document.
- We conclude with Chapter 8 where we recap the three main components of this thesis, comment on the methodological limitations and discuss potential extensions for the future work.

Chapter 2

Related Work

NLI is a type of authorship attribution, which is often approached using a text classification paradigm. We will begin with a discussion of the related problems of text classification and authorship attribution. This will be followed by a complete and detailed review of all Native Language Identification research to date. Finally, we look at Cross-Linguistic Influence, which relates to the language transfer theories that explain the characteristic second language production patterns of learners, an idea that underpins NLI and enables the identification of non-native writers.

Although there has been a significant amount of NLI work presented recently, a survey of the field has yet to be conducted. While this chapter provides pertinent background information for the upcoming chapters, it also serves as the first comprehensive review of NLI research, linking it with wider research areas in linguistics and NLP. In surveying the field, this chapter details the features and experimental setup that will be used in the thesis.

Chapter Contents

2.1 Document Classification

We begin by briefly introducing the concept and methodology of text classification. This is because most of the work in NLI to date has treated the problem as a classification task and tackled it supervised machine learning methods, therefore a brief introduction to the key concepts and problems is warranted.

Document classification (also known as document or text categorization) is the task of automatically assigning text documents to groups using algorithmic methods. These groups may be a set of predefined classes, or in the absence of a pre-existing categories, the algorithm may need to automatically cluster them according to their similarity. This assignment is generally done on the basis of the document's content. Text documents may be classified according to their topic or subject, or some other criteria such as author, document type, writing style or other attributes of the contents. Associated metadata, where available, may also be used in the classification process.

The concept of text classification is not a new one, and its roots can be traced as far back as the work of Maron (1961). The task, which is an amalgamation of the methodologies from Information Retrieval (IR) and Machine Learning (ML), has gained significant attention over the last two decades due to the emergence of a large amount of digital documents and a need to automatically sort and categorize them.

The development of effective ML algorithms has enabled researchers to build automatic document classifiers that learn the attributes and characteristics of interest from a set of labelled texts. Prior to the emergence of these methods documents were manually classified by trained specialists or by systems that used manually engineered classification rules, both of which were cumbersome and error-prone processes. In many cases, given enough training data, machine leaning-based systems can achieve performance comparable to human experts using considerably less time and effort.

2.1.1 The Machine Learning Approach

The ML-based approach to text classification has become the de facto standard over the last two decades. Its high levels of effectiveness and efficiency have established it as a feasible and practical alternative to both manual classification and rule-based classifiers. A very detailed review of this approach is presented by Sebastiani (2002).

The ML approach generally uses a learner which builds a classifier to recognize each category through a general inductive process that involves inspecting a set of *features* extracted from the documents. In this paradigm, most of the effort goes into creating an automatic classifier instead of engineering the classification rules. Although there is a cost associated with the creation of annotated data, this is still often lower than the effort of creating manual classification schemes. The general framework of such a text classification system is shown in Figure 2.1.

This classification procedure may be *unsupervised* or *supervised*. *Unsupervised* classification (Ko and Seo, 2000; Sandler, 2005) is done without any labelled training data, and can be considered a type of document *clustering*. This approach is motivated by the fact that while the collection of training documents is often straightforward, their manual labelling by domain experts is a costly and time consuming process. While some unsupervised systems may attain comparable performance to supervised classifiers on some tasks, their accuracy is generally lower. Such systems can also be used a preliminary step in creating the training data for a supervised classifier.

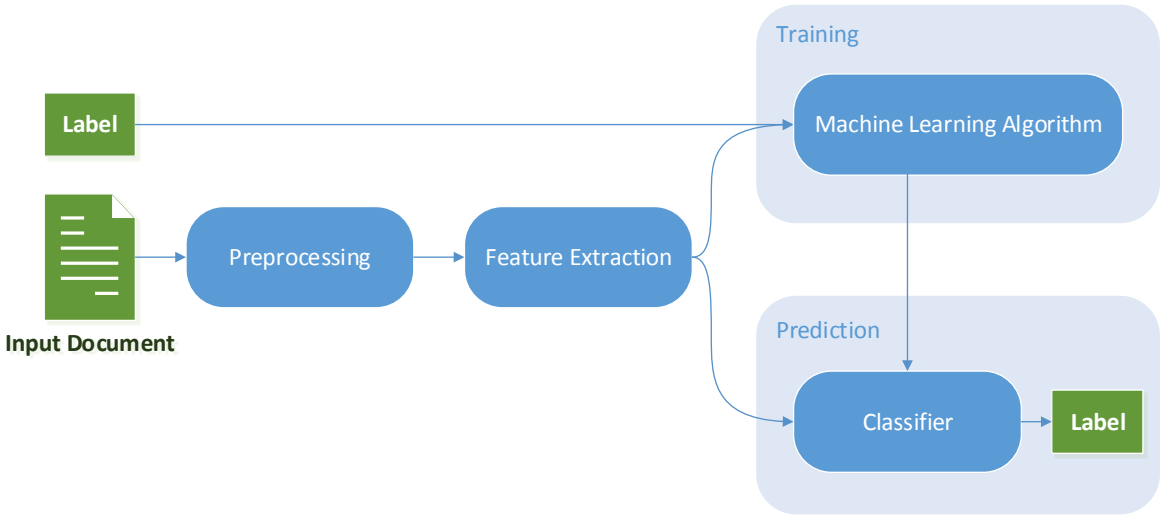

Figure 2.1: General framework for a text classification system.

Supervised classification uses statistical and ML methods to automatically learn classification rules from gold-standard training data that has been manually labelled by human experts. Its name is derived from the fact that the process is *supervised* by the human-assigned categorical information assigned to each training text. In this sense, the training data is a key component of the system. If such pre-labelled data is not available for a classification task, it will need to be created. While a cumbersome and time-consuming task, it is still often easier than manually engineering and maintaining a set of classification rules.

2.1.2 Applications of Text Classification

Document classification techniques can be applied to a variety of problems, as we outline in this section. These problems generally require large amounts of free text documents to be sorted and organized into some set of predefined categories. The types of documents may vary across tasks (*e.g.* news articles, essays, emails, HTML or even tweets), and the type of classification may require input documents to be assigned to a single class or multiple classes. Text classification has been successfully applied to a range of real-world problems and we will briefly review some of them below.

Spam Filtering In this application, incoming email messages are assigned one of two classes: *spam* or *not spam*. This is done by analysing the email contents and the associated metadata such as the sender details.

Document Organization Classification methods can also be applied to organizing a set of documents to facilitate processing and browsing. Examples of such applications include the categorization of news articles or scientific papers into appropriate categories. In more complex scenarios this organization can also be hierarchical, particularly when there are a large number of categories which can be naturally arranged in a hierarchical manner. A good example of such a scenario is the categorization of web pages which are large in number and span hundreds or thousands of categories. Such problems are decomposed into many decisions, each usually at the branch nodes of the hierarchy.

Language Identification Not to be confused with NLI, Language Identification is the task of identifying the language in which a given document has been written in (Cavnar and Trenkle, 1994). In practice this process is performed through learning the association between character distribution models for different languages. Other approaches have also incorporated the use of characteristic words. Such techniques have applications in various document processing systems, including as a first step in automated machine translation (Beesley, 1988). For example, the Google Translate[1] machine translation system currently incorporates such a step if input is entered without explicitly specifying the input language. The training data are usually labelled texts from each of the target languages. Problems can arise when an input document is in a previously unseen language, is too short or contains text from multiple languages. Dialect Identification is a very closely related task, but applied to variants of the same language (Malmasi and Dras, 2015b; Malmasi et al., 2015b).

Sentiment Analysis Opinion mining and sentiment analysis have also made use of text classification techniques. These areas broadly work on the computational treatment of sentiments, opinions, and subjectivity in text that can be used to develop opinion-oriented IR systems. Just as some of the previously mentioned applications need to be topic-aware, these systems must be sentiment-aware and this is often achieved through sentiment classification where the class labels are sentiments such as *positive*, *negative* or *neutral*. The number of categories can vary: sentiment-polarity classification is a binary classification problem but more complex tasks may seek to classify the strength of an opinion with a larger number of classes (*e.g.* determining a star rating for movie reviews). A comprehensive review of the task and the classification approach can be found in the survey by Pang and Lee (2008).

Authorship Attribution The task of identifying the author of a text from a set of candidates is referred to as *authorship attribution*. This task has also been addressed as a text classification problem where the author identities serve as class labels and documents written by those authors are used for training. This task will be discussed in more detail in §2.2.

Genre Classification The classification of a text's genre or style is another problem that has been framed in the text classification context (Lee and Myaeng, 2002; Kessler et al., 1997). Some examples of document genres include research article, product review, news article, legal document, advertisement or novel. It is important to note that these classes are different to the ones used in topic-based classification: genre classification is more concerned with *structure* rather than the content and is often performed with different feature types. It has applications in information retrieval and extraction.

Word Sense Disambiguation The task of identifying the particular sense of an ambiguous (homonymous or polysemous) word is known as *word sense disambiguation* (WSD). In such a task a decision is required about the correct word sense, given the context in which the word appears in. WSD has been approached as a text classification problem where the occurrence contexts are documents and the senses serve as categories (Escudero et al., 2000).

Automated Essay Grading In the educational context, document classification methods have also been adapted to assist in the development of automated grading of student essays (Yannakoudakis

[1] `https://translate.google.com/`

et al., 2011; Shermis and Burstein, 2003; Larkey, 1998). This is motivated by the fact that grading is a highly time consuming process and automated assessment tools could help free up the time of human assessors. It could also help reduce the subjectiveness in grading which leads to significant variance in the marks given by teachers. Such systems have so far used features that rely on the surface forms of the words and the text stylistics; one consequent limitation is that they have generally been unable to assess the *content* being communicated by the essay. However, the results from many systems based on document structure, word use, syntactic fluency and other attributes that measure writing quality have been found to correlate highly with human graders.

Text Filtering Text classification techniques have also been applied to the task of filtering a stream of documents (Belkin and Croft, 1992). Such a system could be used by an information consumer to prevent the delivery of uninteresting incoming documents from an information producer. Alternatively, it could be used on the producer's end to selectively disseminate information to the appropriate consumers, for example in web-based systems that allow uses to create personalized news streams.

2.1.3 Learning Algorithms and Features

Various types of machine learning algorithm have been applied for text classification tasks. These include Naïve Bayes, decision trees, logistic regression and neural networks. Sebastiani (2002) presents a review of how these algorithms are applied in text classification.

Due to the complex nature of human language and texts, the data we work with is often highly dimensional and sparse, requiring robust algorithms that can adapt to such large amounts of data. Many of these learning algorithms cannot scale up to the number of features present in practical text classification problems and thus require dimensionality reduction of the feature space prior to training.

Support Vector Machine (SVM) classifiers have proven to be suited to large feature spaces. They have been widely explored for text classification and categorization and proven to be efficient and effective (Joachims, 1998, 1999). Advantages of SVMs include less reliance on feature reduction as the algorithm is robust against overfitting and can scale up to very large dimensions. Another advantage is that numerous implementations of SVM classifiers are freely and publicly available as a result of the efforts of the research community. These factors have enabled SVMs to provide state-of-the-art performance in many text classification tasks and have been used abundantly in NLI.

Logistic Regression (LR) classifiers, a discriminative model like the SVM, have also proven to be effective for text classification (Genkin et al., 2007) and their use in NLI has also been common. They are linear and probabilistic in nature, enabling them to effectively handle large feature spaces.

A wide range of feature types – *e.g.* word and character n-grams – have been used by researchers to solve various text classification problems. These standard feature classes, along with others used specifically for NLI, are described in detail later in this chapter in §2.3.2.

2.1.4 Evaluation Metrics

The evaluation of text classifiers is generally performed through the analysis of experimental results. These results can be presented within a confusion matrix, which is a table that displays the actual

and predicted classifications provided by the system. Rows contain the actual number of items in each class while the columns display the predicted items in each class.

		Predicted				
	Classes	**1**	**2**	**3**	**4**	**5**
Actual	**1**	[**249**]	0	0	0	121
	2	0	[**87**]	5	0	2
	3	0	19	[**104**]	0	0
	4	32	14	243	[**0**]	26
	5	97	42	64	4	[**12**]

Table 2.1: An example confusion matrix for a classification task with 5 classes. Entries in square brackets along the diagonal represent correct classification.

Each classification result within the possible classes can be qualified as being:

- True Positive (tp): The item has been classified correctly as belonging to the right class.
- False Positive (fp): The item was misclassified as belonging to an incorrect class.
- True Negative (tn): The item was not part of the class and correctly assigned to another class.
- False Negative (fn): The item was part of the class, but incorrectly misclassified.

Based on these criteria, several standard measures have been proposed to measure, evaluate and compare classification performance.

Precision and Recall Originally used in Information Retrieval, Precision and Recall have also been adopted for evaluating document classification.

Precision is defined as the probability that a document predicted to be part of a class actually belongs to that class.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall is defined as the probability that a document that is in a class is predicted to be in this class.

$$\text{Recall} = \frac{tp}{tp + fn}$$

F-measure Combining both precision and recall into a single metric, the F-measure (sometimes called the F_1 measure when it is *balanced*, *viz.* precision and recall are weighted equally) is the harmonic mean of the precision and recall values:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Accuracy Another alternative measure that is commonly used is accuracy. It is a general measure of how well the system has performed and represents the proportion of the results that are true (both positive and negative). Accuracy is influenced by the class distribution.

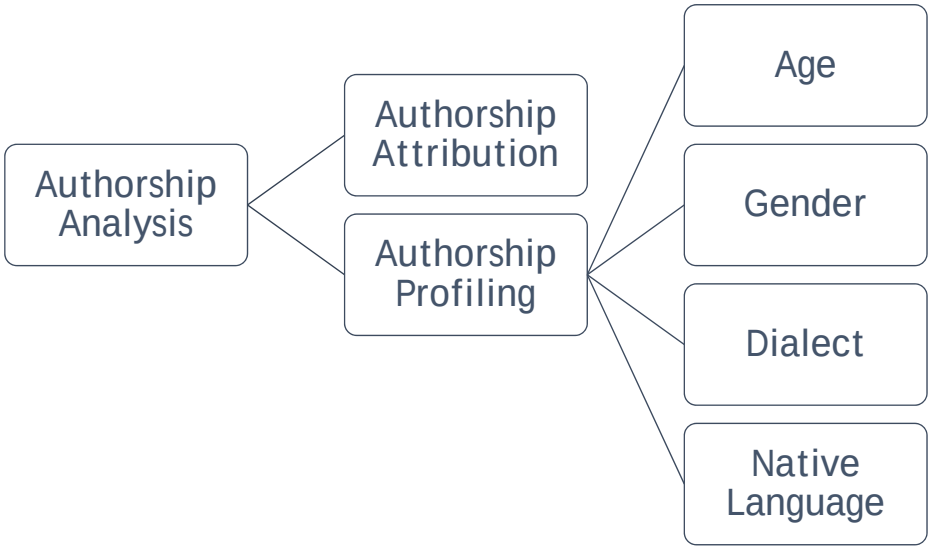

Figure 2.2: Authorship analysis has two major sub-fields. Authorship attribution is concerned with identifying the author from a predefined set of candidate authors. Authorship profiling aims to predict more general traits or characteristics of the author.

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

Other similar metrics for measuring classifier effectiveness have also been proposed, but the ones described here are those that are generally used in the literature.

2.2 Authorship Analysis

The process of analysing the distinct characteristics of a text with the aim of deriving information and conclusions about its author is known as *Authorship Analysis*. This stylometric approach uses statistical methods to analyse the literary style and markers in texts (Holmes, 1998). This area can be broadly represented as two sub-fields: authorship attribution and authorship profiling, as illustrated in Figure 2.2.

Authorship attribution (also known as *authorship identification*) is a well-researched problem with the aim of determining the author of an anonymous text from a predefined list of candidates. This is generally achieved by comparing the input data with previously written texts by the known authors to identify the most likely candidate author. Following the seminal work of Mosteller and Wallace (1964) that used various statistical methods to identify the authors of the Federalist Papers,[2] the techniques have been expanded and applied to various other domains such as scientific writing and social media. The advent of the Internet and social media has been a driving factor in recent authorship attribution research, fuelled by the the rapid growth of user-generated content and publicly available communications and text information. More recently, it has been applied to investigating digital evidence (Chaski, 2005) and analysing extremist communications (Abbasi and Chen, 2005).

[2] The Federalist Papers are an influential collection of documents promoting the ratification of the United States Constitution, published in 1787–88. Although the 85 papers were published under a single pseudonym at the time, the three authors were later identified as being Alexander Hamilton, James Madison, and John Jay. However, the authorship of some of the essays remained disputed and the attribution task here has been to determine which author wrote which of the disputed papers.

Authorship profiling (also known as *authorship characterisation*) aims to predict more general traits or characteristics of the author from their writing. While this is useful in itself, it can also aid authorship attribution investigations where the analysis is approached under the assumption that no viable set of candidate authors is present. Instead, linguistic observations from the text are used to predict characteristics or features that inform us about a particular class that the author belongs to (Argamon et al., 2006). This distinction of authors into several classes enables the creation of author *profiles*, which can then assist with the identification of an author. Various categorical classes of author attributes have been proposed in past research, including age, gender, nationality, socio-economic status, education level and psychological constructs. Estival et al. (2007) investigated several traits and categorized them as being either demographic or psychometric in nature. While many such potential traits have been defined for the purpose of this task, most studies aim to identify a few of these. New traits continue to be proposed in recent studies, including the author's ethnic background as well as whether they are a native English speaker (Rao et al., 2011; Bergsma et al., 2012).

It is also worth mentioning the related area of *similarity detection* methods which can be used to determine the degree of similarity between multiple texts. In this approach, no specific information about the authorship of the works may be required. Similarly, the derived conclusions may also not inform us about the authorship, but rather only the similarity between texts. It could also be employed in cases where the number of potential authors is too large or we cannot be confident that the actual author is within the candidate set (Koppel et al., 2011b). These techniques have also been applied to the problem of automated plagiarism detection.

Chronologically speaking, we can consider authorship analysis to be the predecessor to NLI, and more specifically, we can observe that the goals of NLI and authorship attribution are similar in many respects. While the latter aims to determine the individual author of some text, NLI only seeks to determine the author's native language. If we consider the native language to be a fundamental author trait, then NLI can be seen as a subtask of authorship profiling. Approaches to both tasks have employed similar linguistic features such as function words and character n-grams.

2.3 Native Language Identification

The process of determining an author's native language based on their language production in a non-native language is known as Native Language Identification. In particular, we can consider this to be a type of authorship profiling, as discussed in §2.2.

NLI works by identifying language use patterns that are common to certain groups of speakers that share the same native language. This process is underpinned by the presupposition that an author's linguistic background will dispose them towards particular language production patterns in their learnt languages, as influenced by their mother tongue. This relates to the issue of *Cross-linguistic Influence*, which will be discussed in §2.4. The diagram in Figure 2.3 illustrates the concept of an example NLI system.

NLI is a fairly recent, but rapidly growing, area of research. While some early research was conducted in the early 2000s, most work has only appeared in the last few years. This trend can be observed in Figure 2.4 which shows the number of NLI related publications per year since the first

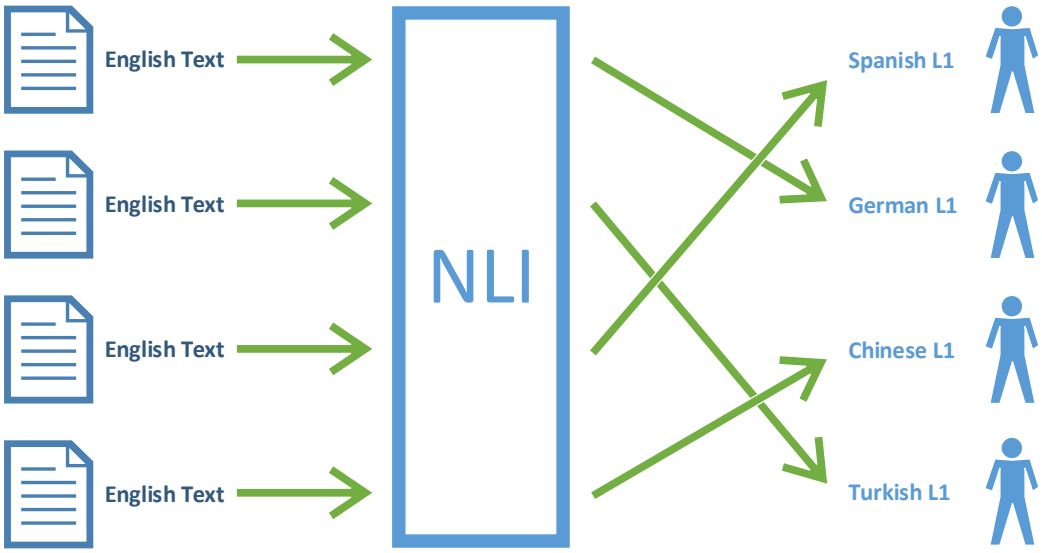

Figure 2.3: The general concept of an NLI system is depicted in this example. English texts that are written by non-native authors are passed to an NLI system which predicts the native language of their authors based on their writing.

study in this area. This surge of interest, coupled with the inaugural shared task in 2013, has resulted in NLI becoming a well-established NLP task.

While NLI can help us better observe and understand the ways in which non-natives learn a language, the methodology can interact with techniques from SLA in analyzing the process of language learning. The results of these kinds of analyses can help learners improve their language skills in a self-directed manner and also assist teachers and educators to better focus their efforts on particular areas that prove difficult for students from specific L1 backgrounds.

In many areas of NLP, the focus of research is on automating tasks that can be performed accurately and without much difficulty by humans, but not machines. Examples include speech recognition, classification and named entity extraction. In contrast, there are other tasks such as authorship attribution and NLI that are quite complex for humans, but much easier to do through statistical analysis and machine learning. It is quite plausible that given the number of possible candidate languages and the large range of feature types and spaces, it would be unlikely that human judges could match the performance of current NLI systems.

NLI is well suited to languages that have many non-native speakers. English, due to its prominence, large number of learners and role as business and scientific *lingua franca* has attracted the most attention from researchers to date. This has also been facilitated by the large body of learner data that has accumulated over the last few decades. With this in mind, one of the central contributions of this work is the extension of NLI to additional, non-English L2s such as Chinese.

In most studies in the field, NLI has been treated as a multi-class supervised classification task. In this common experimental design, the native languages (L1) are used as class labels and the individual writings are used as training and testing data. Results are usually reported in terms of the classification accuracy, often under cross-validation.

NLI systems can be improved and optimized in several ways, including:

- Classification algorithm selection
- Hyperparameter optimization

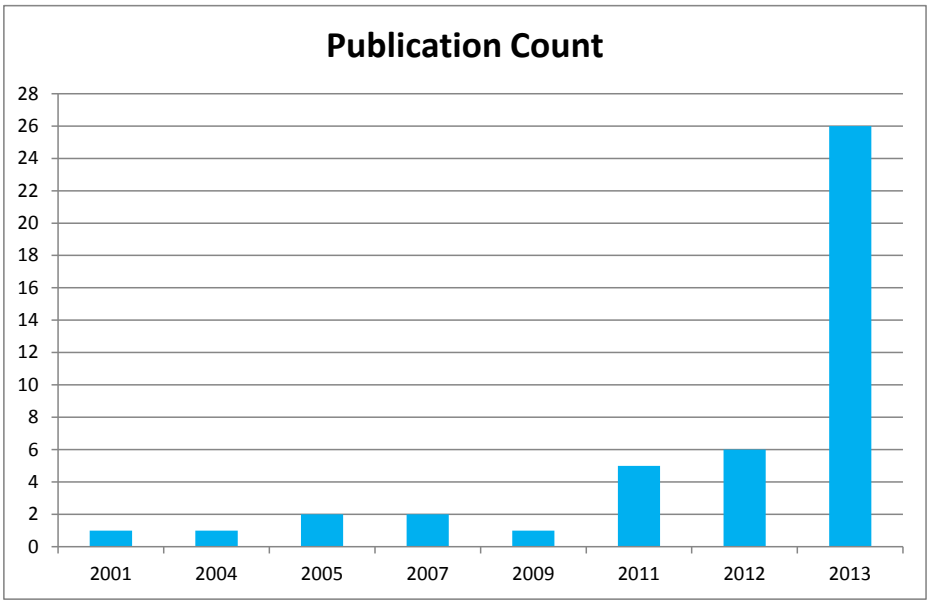

Figure 2.4: The number of NLI related scientific publications per year, for the years 2001–2013.

- Incorporation of useful feature types and spaces
- Feature selection and representation optimization
- Feature combination methods

Most NLI research works on improving some or all of these facets of the task.

2.3.1 NLI Corpora

A number of different datasets have been used for NLI. In this section we introduce some of the most commonly used ones.

Learner corpora — datasets composed of the writings of learners of a particular language — are a key component of language acquisition research and their utilization has been considered "a revolution in applied linguistics" (Granger, 1994).

They are designed to assist researchers studying various aspects of learner interlanguage and are often used to investigate learner language production in an exploratory manner in order to generate hypotheses. Recently, learner corpora have also been utilized in various NLP tasks including error detection and correction (Gamon et al., 2013), language transfer hypothesis formulation (Swanson and Charniak, 2014) and Native Language Identification (Tetreault et al., 2013). In fact, they are a core component of NLI research. While such corpus-based studies have become an accepted standard in SLA research and relevant NLP tasks, there remains a paucity of large-scale L2 corpora.

Additionally, there are a number of characteristics and design requirements that must be met for a corpus to be useful for NLI research. An ideal NLI corpus should:

- have multiple and diverse L1 groups represented
- be balanced by topic so as to avoid topic bias, as discussed in the next section
- be balanced in proficiency across the groups
- contain similar numbers of texts per L1, *i.e.* be balanced by class

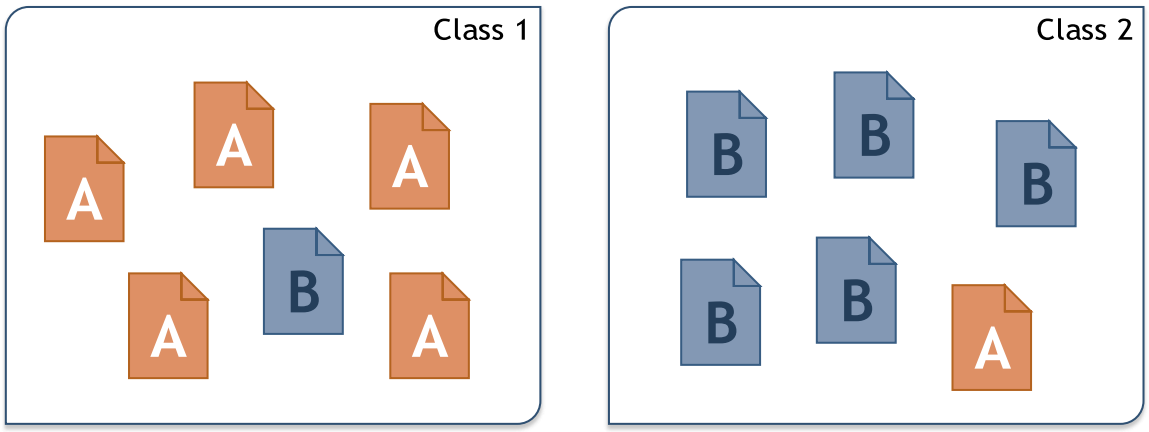

Figure 2.5: An example of a dataset that is not balanced by topic: class 1 contains mostly documents from topic *A*, while class 2 is dominated by texts from topic *B*. Here, a learning algorithm may distinguish the classes through other confounding variables related to topic.

- be sufficiently large in size to reliably identify inter-group differences

Unfortunately very few of the currently available data sources meet all of these criteria.

2.3.1.1 Topic Bias

Before we go on to discuss the available corpora, we first introduce the concept of *topic bias*, an important criteria in evaluating these datasets. Topic bias, an issue that affects stylistic classification tasks like NLI, can occur as a result of the themes or topics of the texts to be classified not being evenly distributed across the classes. For example, if in our training data all the texts written by English L1 speakers are on topic A, while all the French L1 authors write about topic B, then we have implicitly trained our classifier on the topics as well. In this case the classifier learns to distinguish our target variable through another confounding variable. This concept is illustrated in Figure 2.5. While this is an extreme example, in real world scenarios even a slightly skewed topic distribution can negatively affect the classification results and thus this source of bias must be accounted for.

Accordingly, the use of lexical features (*e.g.* word *n*-grams) cannot be justified in circumstances where a topic balanced dataset is not being used, due to the mentioned issues with topic bias (Koppel et al., 2009). For this reason, some of the earlier studies we mention in our review of the literature did not use lexical features when investigating unbalanced datasets such as the ICLE. Other researchers like Brooke and Hirst (2012b), however, argue that lexical features cannot be simply ignored.

2.3.1.2 The ICLE Corpus

The *International Corpus of Learner English* (ICLE) contains English essays written by learners from 16 different native language backgrounds (Granger, 2003). These L1 groups and the number of essays in each group are shown in Table 2.2.

It should be noted that this dataset was not designed for use in NLI research and researchers have pointed out a number of methodological flaws that may arise due to its use. A major issue is that the topic distribution is not balanced across the L1 groups. Brooke and Hirst (2011) showed that the corpus is compromised by topic bias that makes it difficult to draw conclusions about the true efficacy of the derived features. These findings are discussed later in this section.

Native Language	Documents
Bulgarian	302
Chinese	982
Czech	243
Dutch	263
Finnish	390
French	347
German	437
Italian	392
Japanese	366
Norwegian	317
Polish	365
Russian	276
Spanish	251
Swedish	355
Tswana	519
Turkish	280
Total	6,085

Table 2.2: The 16 L1 groups included in the ICLE corpus and the number of essays in each group.

Native Language	Documents
Catalan	64
Chinese	66
Dutch	2
French	146
German	69
Greek	74
Italian	76
Japanese	81
Korean	86
Polish	76
Portuguese	68
Russian	83
Spanish	200
Swedish	15
Thai	63
Turkish	75
Total	1,244

Table 2.3: The 16 L1 groups included in the CLC FCE corpus and the number of essays in each group.

Another issue with this corpus is that it is quite small, containing only 6,085 texts, making it difficult to effectively split the data for training and testing. Any additional splitting of the already limited texts limits our ability to obtain the statistically significant results. Furthermore, it makes it difficult to determine whether the developed methodology can scale well to larger corpora or be applied to different domains.

Prior to the 2013 NLI Shared Task, most researchers had used the second version of the ICLE corpus for training and evaluation as the best available option at that time.

2.3.1.3 The CLC FCE Corpus

The CLC FCE Dataset[3] is a set of 1,244 exam scripts written by candidates who sat the Cambridge ESOL First Certificate in English (FCE) examination in 2000 and 2001. The scripts are extracted

[3]http://ilexir.co.uk/applications/clc-fce-dataset/

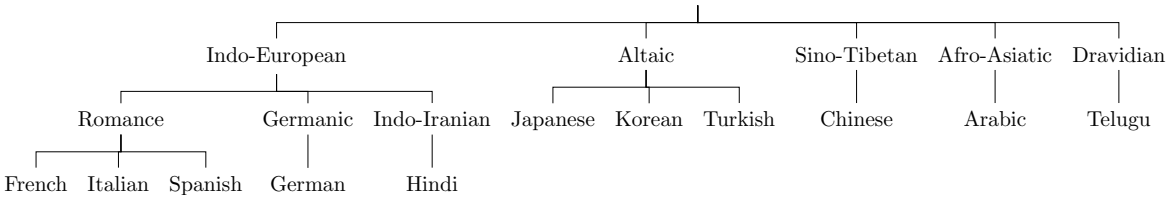

Figure 2.6: Language families in the TOEFL11 corpus. The languages were selected to represent different families, but to also have several from within the same families. Diagram reproduced from Blanchard et al. (2013).

from the much larger Cambridge Learner Corpus (CLC),[4] which is being developed as a collaborative effort between Cambridge University Press and Cambridge Assessment.

The 1,244 scripts are distributed across 16 L1 groups, as shown in Table 2.3. For each exam script, the CLC FCE Dataset includes the original text written by the candidate (transcribed and anonymised, but otherwise unmodified) as well as marks, error annotation and essential demographic details including the candidate's first language and age bracket.

Although this corpus has been used for some NLI research, it is considered too small for this line of research as the majority of the L1 groups have fewer than 80 texts.

2.3.1.4 The TOEFL11 Corpus

The TOEFL11 corpus (Blanchard et al., 2013) — also known as the *ETS Corpus of Non-Native Written English* — is the first dataset designed specifically for the task of NLI and developed with the aim of addressing the above-mentioned deficiencies of other previously used corpora. By providing a common set of L1s and evaluation standards, the authors set out to facilitate the direct comparison of approaches and methodologies.

Furthermore, as all of the texts were collected through the Educational Testing Service's electronic test delivery system, this ensures that all of the data files are encoded and stored in a consistent manner.[5] The corpus is available through the the Linguistic Data Consortium.[6]

It consists of 12,100 learner texts from speakers of 11 different languages. The texts are independent task essays written in response to eight different prompts, and were collected in the process of administering the Test of English as a Foreign Language (TOEFL®) between 2006-2007. The texts are divided into specific training (TOEFL11-TRAIN), development (TOEFL11-DEV) and test (TOEFL11-TEST) sets.

The 11 L1s are Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu and Turkish. This selection ensures that there are L1s from diverse language families, but also several from within certain families. The L1s and their language families are shown in Figure 2.6.

This dataset was designed specifically for NLI and the authors attempted to balance the texts by topic and native language. There are a total of eight essay prompts in the corpus, with the prompts setting each essay's topic or theme. Although they were not able to create a perfectly balanced corpus, the distribution of topics across L1s is very even. This distribution of essay prompts by L1 is shown in Figure 2.7. Additionally, the texts of the eight prompts are listed in Table 2.4.

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

[5] The essays are distributed as UTF-8 encoded text files.

[6] https://catalog.ldc.upenn.edu/LDC2014T06

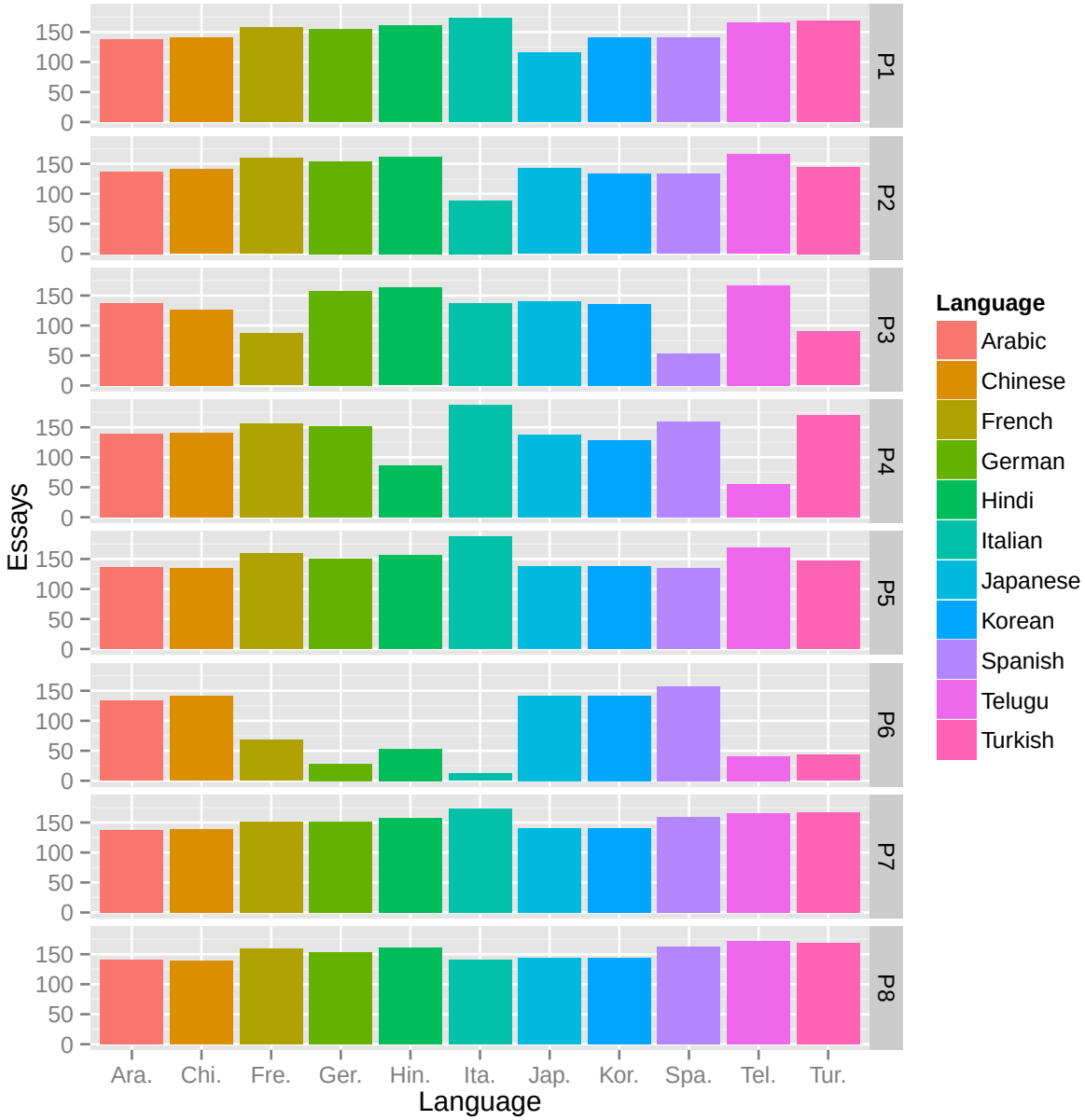

Figure 2.7: A plot of how the eight topic prompts are distributed across the L1 groups in the TOEFL11 corpus. Prompts are labelled P1–P8. Figure reproduced from Blanchard et al. (2013).

Prompt	Text
P1	Do you agree or disagree with the following statement? It is better to have broad knowledge of many academic subjects than to specialize in one specific subject. Use specific reasons and examples to support your answer.
P2	Do you agree or disagree with the following statement? Young people enjoy life more than older people do. Use specific reasons and examples to support your answer.
P3	Do you agree or disagree with the following statement? Young people nowadays do not give enough time to helping their communities. Use specific reasons and examples to support your answer.
P4	Do you agree or disagree with the following statement? Most advertisements make products seem much better than they really are. Use specific reasons and examples to support your answer.
P5	Do you agree or disagree with the following statement? In twenty years, there will be fewer cars in use than there are today. Use reasons and examples to support your answer.
P6	Do you agree or disagree with the following statement? The best way to travel is in a group led by a tour guide. Use reasons and examples to support your answer.
P7	Do you agree or disagree with the following statement? It is more important for students to understand ideas and concepts than it is for them to learn facts. Use reasons and examples to support your answer.
P8	Do you agree or disagree with the following statement? Successful people try new things and take risks rather than only doing what they already know how to do well. Use reasons and examples to support your answer.

Table 2.4: Texts of the eight prompts in the TOEFL11 corpus.

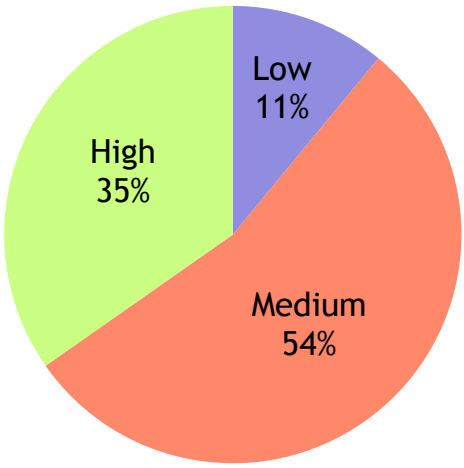

Figure 2.8: Distribution of proficiency levels in the TOEFL11 corpus.

Furthermore, the proficiency level of the author of each text (low, medium or high) is also provided as metadata. More specifically, there are 1,330 low, 6,568 medium and 4,202 high proficiency texts. This distribution is visualized in Figure 2.8. It should also be noted that the proficiency levels are not evenly distributed across languages. Some languages such as Arabic, Japanese and Korean have higher numbers of low proficiency essays and fewer high proficiency ones. The complete distribution of proficiency levels per L1 in the TOEFL11 corpus is illustrated in Figure 2.9.

2.3.2 Classification Features for NLI

A wide range of feature types have been used by researchers to solve various text classification problems, including NLI. They vary in their linguistic complexity, ranging from shallow surface-based features to those using deeper linguistic information. The nature of these features can be broadly

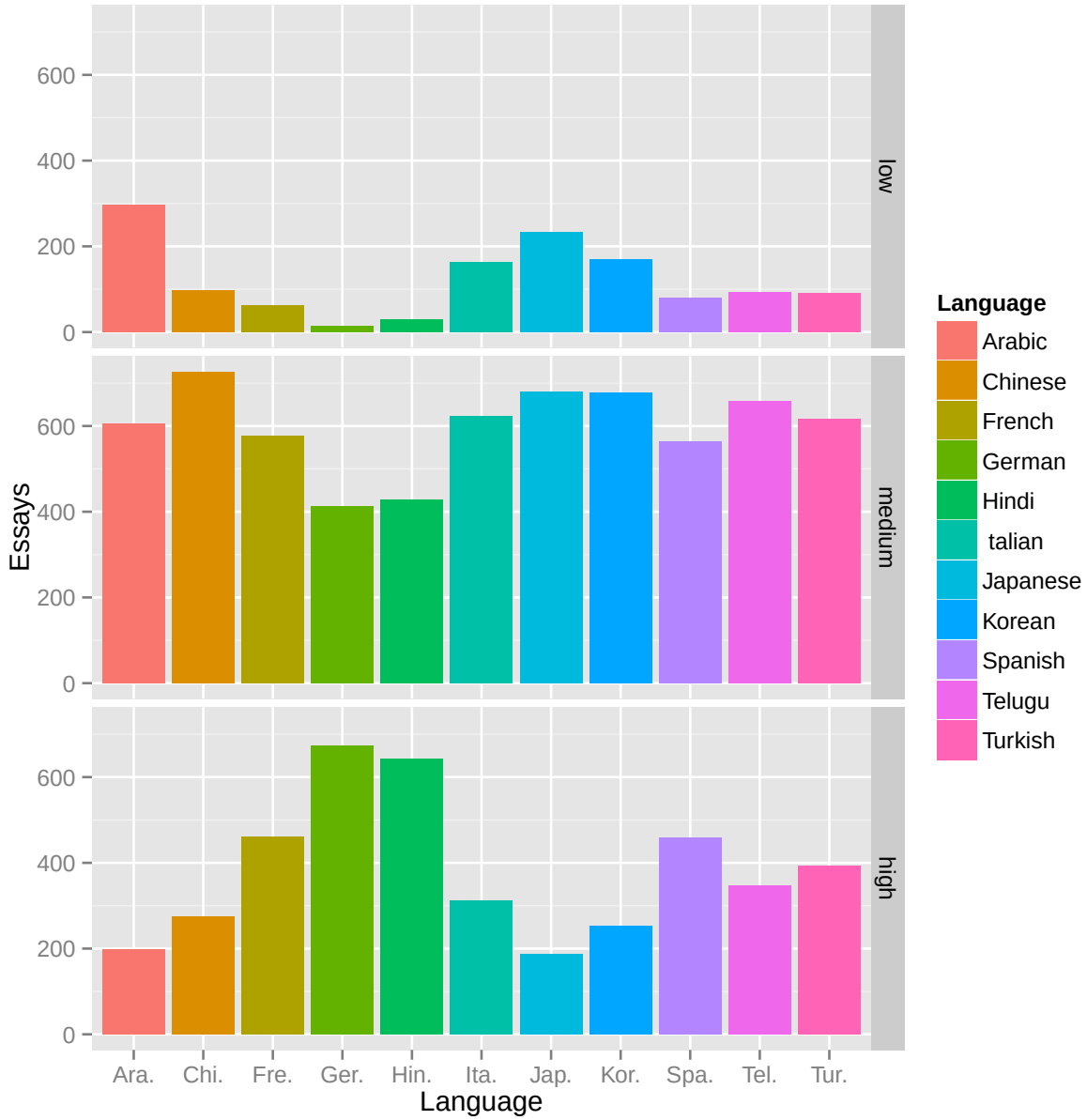

Figure 2.9: The complete distribution of proficiency levels per L1 in the TOEFL11 corpus. Figure reproduced from Blanchard et al. (2013)

categorized as being lexical, syntactic or stylistic. In this section we will discuss how these features are extracted and describe a number of these features that have been used for text classification.

While using the aforementioned lexical features allows for a shallow analysis of the text by mostly using the surface forms of the words, a deeper and more linguistically sophisticated approach can be taken using syntactic features.

Many of these features are associated with specific linguistic cues that can assist with the language identification. Cross-linguistic transfer has been shown to be present within the various linguistic subsystems such as phonology, syntax, morphology and lexicon. Many of these features attempt to capture discriminative cues that are related to these linguistic systems as expressed through an individual's writing.

2.3.2.1 *N*-gram frequency models

The use of n-gram frequency models is very common in text classification tasks and this approach generally involved the creation of n-gram frequency profiles for document classes and comparing them (Cavnar and Trenkle, 1994).

An *n-gram* can be defined as a sub-sequence of n contiguous items taken from a larger sequence. The items in the sequence may be characters, words or other sequential data such as part of speech tags, phonemes or syllables.

For most data types, not all items have the same probability of occurring, some will inevitably appear more frequently than others and these distributions often vary by language, domain or register. For example, this principle is embodied by Zipf's Law (Zipf, 1949) which when applied to English vocabulary reveals that word usage frequency follows a type of power-law distribution. That is to say, while many words are used, only a few words can account for most of usage and the rest are rarely used.

By the same token, words, characters and other items do not combine together in a random or equally likely way. In fact, analysing an item and its neighbours as a sequence of n items can be highly informative. Such sequences exist in various aspects of language use (*e.g.* due to grammatical constraints or lexical associations) and can be exploited for classifying documents, particularly by using distributions of n-grams. The use of distributions also enables systems to be robust against statistical outliers. For text data, these may be a result of human errors (*e.g.* spelling mistakes) or part of the digitization process (*e.g.* Optical Character Recognition errors in scanning).

To use such distributions, n-grams of a given feature type are first extracted in each document class and the occurrences of each n-gram feature are counted to generate a frequency profile. When a document is to be classified, this frequency distribution is obtained by the same procedure and the extracted feature n-grams can be compared against those of each category to find the best match. One issue to keep in mind is that suitable training data is generally needed to generate representative frequency distributions. The training data must have a reasonable number of documents and each document must be sufficiently long. Using a few documents or very short texts may lead to unrepresentative statistics.

Today n-gram models are widely used in broad range of applications, including language identification, topic classification and speech recognition. They have been found to be a simple and reliable method of using textual features and will be used in conjunction with many of the text classification features that we describe below.

It is also important to note that these models can become quite large, and consequently sparse, as the value of n increases. Theoretically, the n-gram space grows exponentially with n, but in practice this growth rate is somewhat lower as not all possible combinations are observed in the data.

2.3.2.2 Character n-grams

This is a sub-word feature that uses the constituent characters that make up the whole text. When used as n-grams, the features are n-character slices of the text. From a linguistic point of view, the substrings captured by this feature, depending on the order, can implicitly capture various sub-lexical features including letters, phonemes, syllables, morphemes and suffixes.

There may be some implementation-specific variation in the method for extracting the character n-grams. They may be extracted from across the whole text or limited to single words (*i.e.* substrings are only extracted from within word boundaries). In other cases word boundaries may be kept and represented by a special character such as an underscore("_"). Furthermore, the letter cases of the words may also be normalized by folding them to all capitals or lowercase before extracting n-grams. These choices can vary by implementation and some choices, such as converting all characters to uppercase instead of lowercase, may not lead to different results.

As an example, for the value of $n = 3$, the character trigrams for the word *language* are $\{lan, ang, ngu, gua, uag, age\}$. To use them for classification, the frequencies of all such n-grams are calculated in each document and passed to the machine learner which uses them to train a classifier model.

2.3.2.3 Word n-grams

The surface forms of words can be used as a feature for classification. The features are extracted by tokenizing the documents. Each unique token may be used as a feature, but the use of n-gram distributions is also common. In this scenario, the n-grams are extracted along with their distributions. The example below demonstrates the word n-grams can be extracted from a sentence.

(2.1) John peered through the door.

The sentence in example 2.1 would generate the following bigrams: {*John peered*, *peered through*, *through the*, *the door*}. The bi-gram feature captures various collocations from the text. Performed over a large amount of text, the bi-gram distribution will help us identify the strongest (*i.e.* most frequently occurring) collocations in the data.

The texts may optionally be converted to all lower or upper case letters in situations where letter case information is not informative for classification. This normalization will help to reduce the number of features in the feature space. Another variation is the exclusion of punctuation tokens.

The extraction of such n-grams of various orders can create a high dimensional feature space containing thousands, if not millions, of such n-gram items. These distributions can be used by machine learners to identify features that are highly discriminative and thus useful for document classification.

2.3.2.4 Lemma n-grams

With this feature, the tokens are first lemmatized to convert them to a normal form. This generally means that the lexemes in a document are replaced with their lemmas. The example below demonstrates how the various inflected forms of word are mapped to a single token.

$$\{run,\ ran,\ runs,\ running\} \rightarrow RUN$$

Some lemmatizers can use context information (such as word part-of-speech) and dictionary information to correctly map polysemous or irregularly inflected words to their correct lemma. Other approaches may simply rely on a mapping of tokens to their lemmas. Both of these approaches are vulnerable to spelling mistakes and rare words which are considered out-of-vocabulary. The use of lemmas can be useful in generalizing the word token information in highly inflectional languages where a word may have many forms.

2.3.2.5 Function Words

In contrast to content words, function words do not refer to specific topics themselves, but rather can be seen as indicating the grammatical relations between other words. In a sense, they are the syntactic glue that hold much of the content words together and their role in assigning syntax to sentences is linguistically well-defined. They generally belong to a language's set of closed-class words and embody relations more than propositional content. Examples include articles, determiners, conjunctions and auxiliary verbs.

Function words are considered to be highly context- and topic-independent but other open-class words can also exhibit such properties. In practical applications, such as Information Retrieval, such words are often removed as they are not informative and stoplists for different languages have been developed for this purpose. These lists contain 'stop words' and formulaic discourse expressions such as *above-mentioned* or *on the other hand*.

Function words' topic independence has led them to be widely used in studies of authorship attribution (Mosteller and Wallace, 1964) as well as NLI[7] and they have been established to be informative for these tasks. Much like Information Retrieval, the function word lists used in these tasks are also often augmented with stoplists and this is also the approach that we take.

Such lists generally contain anywhere from 50 to several hundred words, depending on the granularity of the list and also the language in question.

2.3.2.6 Part-of-Speech Tags

Parts of Speech are linguistic categories (or word classes) assigned to words that signify their syntactic role. Basic categories include verbs, nouns and adjectives but these can be expanded to include additional morpho-syntactic information. The assignment of such categories to words in a text enables us to add a level of linguistic abstraction by marking each word with its appropriate POS.

In computational linguistics this task is generally done automatically using a POS Tagger that has been designed for a specific language. This task is often complicated by lexical ambiguity, which refers to the fact that the same lexical representation or surface form of a word may have different

[7] For example, the largest list used by Wong and Dras (2009a) was a stopword list from Information Retrieval; given the size of their list, this was presumably also the case for Koppel et al. (2005a), although the source there was not given.

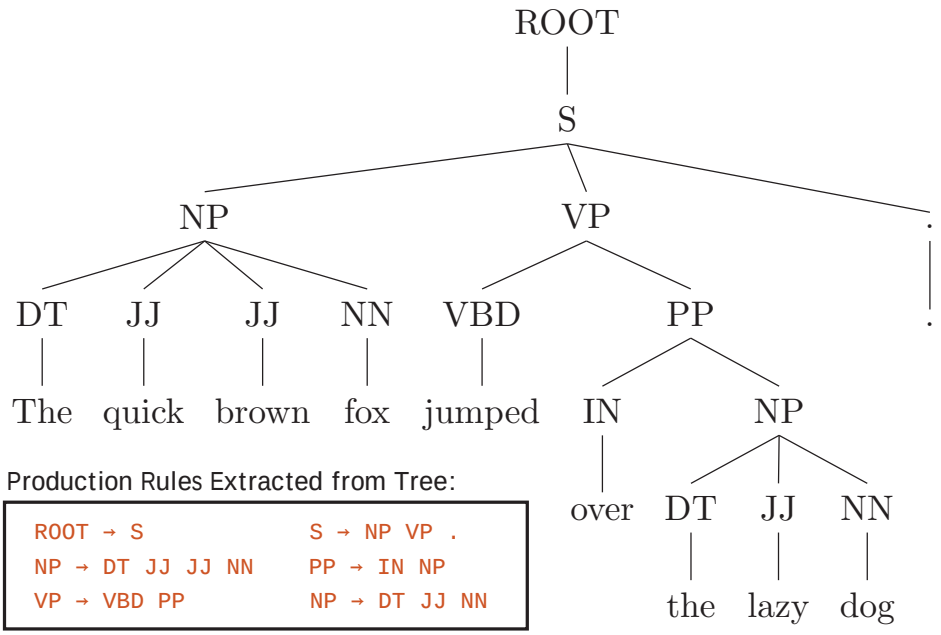

Figure 2.10: A constituent parse tree for an example sentence along with the context-free grammar production rules which can be extracted from it.

meanings, depending on the context. To illustrate, compare the use of the word *walks* in the following two sentences:

(2.2) John *walks* by the beach every Wednesday.

(2.3) John enjoys *walks* by the beach each Wednesday.

POS taggers usually disambiguate the word classes by taking into account the context and relationship with adjacent words. More thorough expositions of such systems can be found in some of the seminal works in this field (Santorini, 1990; Brill, 1992; Schmid, 1994; Ratnaparkhi, 1996).

For text classification, analysis of their n-gram distributions can enable a classifier to recognize linguistic regularities. POS n-grams capture the syntactic patterns, constructions and structure of the texts and have been proven to be useful in classification tasks such as authorship attribution.

2.3.2.7 Phrase Structure Rules

Also known as Context-free Grammar Production Rules, these are the rules used to generate constituent parts of sentences, such as noun phrases. One way to obtain these is by first generating constituent parses for all sentences. The production rules, excluding lexicalizations, are then extracted. Figure 2.10 illustrates this with an example tree and its rules.

These context-free phrase structure rules capture the overall structure of grammatical constructions and global syntactic patterns. They can also encode highly idiosyncratic constructions that are particular to some L1 group. They have been found to be useful for NLI (Wong and Dras, 2011).

2.3.2.8 Tree Substitution Grammar Fragments

Tree Substitution Grammar (TSG) fragments have been proposed by Swanson and Charniak (2012) as yet another type of syntactic feature for NLI or other syntactically motivated text classification tasks. They demonstrated that this feature type can achieve high classification accuracy.

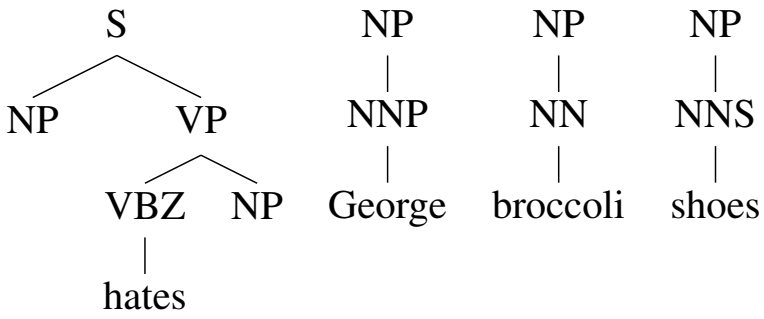

Figure 2.11: Fragments from a Tree Substitution Grammar capable of deriving the sentences "George hates broccoli" and "George hates shoes". Reproduced from Swanson and Charniak (2012).

TSGs are a generalization of context-free grammars that allow non-terminals to rewrite as fragments which can have an arbitrary size (Post and Gildea, 2013), instead of being limited to a depth of one. A TSG *fragment* or *elementary tree* refers to these rules. Figure 2.11 shows several example fragments from a Tree Substitution Grammar capable of deriving the sentences "George hates broccoli" and "George hates shoes".

2.3.2.9 Grammatical Dependencies

A dependency grammar captures the grammatical relationship between words in a sentence, usually a relation between a head and its dependents.

In Tetreault et al. (2012), Stanford dependencies were investigated as yet another form of syntactic feature. To extract this feature, each sentence in the texts must be parsed with a dependency parser to obtain the grammatical dependencies. For the sentence shown in Figure 2.10 we obtain the following dependencies:

```
det(fox-4, The-1)
amod(fox-4, quick-2)
amod(fox-4, brown-3)
nsubj(jumps-5, fox-4)
root(ROOT-0, jumps-5)
case(dog-9, over-6)
det(dog-9, the-7)
amod(dog-9, lazy-8)
nmod(jumps-5, dog-9)
```

The frequencies of each one of these dependency relations (without the numeric index) are then counted across the whole text and the frequency distribution of all these relations are used as features, similar to other n-gram features.

The dependencies have several styles of representation. In *basic* dependencies — the representation used by Tetreault et al. (2012) — each word is a dependent of one other word. A *collapsed* representation converts the tree structure into a directed graph and thus considers additional dependencies such as those from relative clauses and their antecedents.

Additionally, a POS transformation can also be applied to generate all the variations for each of the dependencies (grammatical relations) by substituting each lemma with its corresponding POS tag. For instance, a grammatical relation of `det(dog, the)` yields the following variations: `det(NN, the)`,

`det(dog, DT)`, and `det(NN, DT)`. This greatly increases the number of features but also generates more abstract relations that can better generalize across documents.

These grammatical dependencies have been found to be a very useful NLI feature and thought to capture a "more abstract representation of syntactic structures" (Tetreault et al., 2012; Bykh and Meurers, 2014).

2.3.2.10 Adaptor Grammar n-grams

Adaptor grammars (Johnson, 2010) are a generalization of probabilistic context-free grammars that can capture collocational pairings. Wong et al. (2012) explore the use of adaptor grammars to discover arbitrary length n-gram collocations in a corpus.

They explore both the pure part-of-speech (POS) n-grams as well as the more promising mixtures of POS and function words. This mixed model retains the surface form of function words instead of using their POS tag. Some example mixed trigrams include "`the NN that`" or "`NN that VBZ`". They demonstrated that such features can outperform their pure POS counterparts.

They derive two adaptor grammars where each is associated with a different set of vocabulary: either pure POS or the mixture of POS and function words. They use the grammar proposed by Johnson (2010) for capturing topical collocations as presented below:

$$
\begin{array}{ll}
Sentence \rightarrow Doc_j & j \in 1, \ldots, m \\
Doc_j \rightarrow \_j & j \in 1, \ldots, m \\
Doc_j \rightarrow Doc_j\ Topic_i & i \in 1, \ldots, t; \\
 & j \in 1, \ldots, m \\
\underline{Topic_i} \rightarrow Words & i \in 1, \ldots, t \\
Words \rightarrow Word & \\
Words \rightarrow Words\ Word & \\
Word \rightarrow w & w \in V_{pos}; \\
 & w \in V_{pos+fw}
\end{array}
$$

As per Wong et al. (2012), V_{pos} contains 119 distinct POS tags based on the Brown tagset and V_{pos+fw} is extended with 398 function words used in Wong and Dras (2011). The inference algorithm for the adaptor grammars are based on the Markov Chain Monte Carlo technique made available by Johnson (2010).[8] Each one of the arbitrary length n-grams and their frequency is then used a classification feature.

2.3.2.11 Summary of Features

A large number of features that have been explored for text classification and NLI and the selection presented here is not exhaustive. There is no absolute best feature and the choice of which features to use depends on the task. In this section we described the most commonly used and notable features, including ones that will be used in this thesis.

[8] `http://web.science.mq.edu.au/~mjohnson/Software.htm`

2.3.3 Related Work

2.3.3.1 Early Studies

To our knowledge, the earliest work on detecting non-native language is that of Tomokiyo and Jones (2001) whose main aim was to detect non-native speech using part-of-speech and lexical features, and to also determine the native language of the non-native speakers. They were able to achieve 100% accuracy in their study, which included six Chinese and 31 Japanese speakers.

Inspired by the field of stylometry, Jarvis et al. (2004) presented an early approach to L1 transfer using lexical style and word choice, referred to as *wordprints*. They compiled a corpus of 446 texts written by adolescent EFL learners from five different L1 backgrounds (Danish, Finnish, Portuguese, Spanish and Swedish). Closely related L1 pairs were selected in order to make the task more challenging. The corpus texts consisted of descriptions of a segment of a silent Charlie Chaplin film. Focusing exclusively on lexical transfer, the 30 most frequent words used by each group were collected, resulting in a set of 53 words which were used as classification features. They applied a Linear Discriminant Analysis (LDA) classifier with 11-fold cross-validation, obtaining an average accuracy of 81%.

Koppel et al. (2005a; 2005b) reported one of the first and most significant results in this field (though many studies have erroneously referred to their work as being the very first in this field, this is not entirely correct). Texts ranging from 500–850 words from five native languages (Bulgarian, Czech, French, Russian, and Spanish, 258 texts per language) were selected from the first version of the International Corpus of Learner English (Granger et al., 2009). They used a set of syntactic, lexical, and stylistic features that included function words (*e.g. and*, *from*, *which*, *whilst*), character n-grams (*i.e.* sequences of characters of length n), and part-of-speech (PoS) bi-grams, together with some spelling mistakes. Using an SVM classifier, the achieved a classification accuracy of 80% with ten-fold cross-validation. This was a strong result, given the chance baseline of 20%. The authors note that just using character n-grams and function words enabled them to achieve an accuracy of 75%, highlighting the utility of those specific features. Koppel et al. also suggested that syntactic features (specifically errors) might be potentially useful, but only explored this idea at a rather shallow level by characterising ungrammatical structures with rare POS bi-grams.

Tsur and Rappoport (2007) also used the same setup as Koppel et al. (2005a) in order to examine the role of phonology in the non-native speaker's lexical choice. The authors hypothesized that character n-grams were useful in capturing the syllables of a language's phonology, which would in turn be influenced by the phonology of the writer's native tongue. Using the same set of five languages as Koppel et al. (2005a), they extracted the 200 most frequent bigrams for use as classification features. Using this method they obtained an accuracy of 66% using the bigrams and compared it to a baseline of 46% that was achieved using character unigrams. In a second experiment they replaced two of the five languages (French and Spanish were replaced with Italian and Dutch) and applied their methodology again, obtaining another similar result of 65%. On the whole, they were able to successfully replicate the earlier results of Koppel et al. (2005a) and demonstrated that it was possible to achieve good results simply by using character n-gram frequency distributions. The claim that such n-grams are related to L1 phonology has since been challenged, for example by Nicolai and Kondrak (2014), as we discuss in §2.3.3.7.

Estival et al. (2007) conducted a similar experiment using a proprietary dataset of 9,836 English emails written by speakers of three languages (English, Spanish and Arabic) with a majority class

baseline of 62.9%. They employed a set of lexical, syntactic and document structure (*i.e.* HTML) features, testing them with several different types of classifier. They report a best accuracy of 84% on this three way classification problem using a Random Forest classifier. It should be noted that these results cannot be directly compared to the previous experiments as they use different numbers of classes.

While the approaches described so far have relied heavily on lexical features, they made little use of the syntactical cues available in the texts. Wong and Dras (2009a) proposed the use of syntactic information for NLI as a domain- and content-independent source of clues for classification, motivated by the initial insights from Contrastive Analysis (CA).[9] Using the five languages investigated by Koppel et al. with the addition of Chinese and Japanese, they explored the utility of syntactic error information for NLI. They first replicated the work of Koppel et al. (2005a) with the three classes of feature mentioned above and then extended the classification model with three syntactic errors commonly observed in non-native English users which in CA had been identified as being influenced by the native language: subject-verb disagreement, noun-number disagreement and misuse of determiners. This preliminary study provided some evidence for the usefulness of syntactic information, supported by statistical ANOVA tests which showed that languages differed significantly in their error distributions. With regards to NLI, however, this did not yield an improvement in classification accuracy.

2.3.3.2 Growing Interest

Following these preliminary studies, there was a spike in interest in NLI over the next few years that led to marked advancements in the methods used to tackle the problem.

Wong and Dras (2011) extended their earlier results by exploiting parse structures for NLI. They explored the usefulness of syntactic features in a broader sense by characterising syntactic errors with cross sections of parse trees obtained from statistical parsing. More specifically, they utilised two types of parse tree substructure to use as classification features — horizontal slices of the trees and the feature schemas used in discriminative parse reranking (Charniak and Johnson, 2005). Only using non-lexicalized rules and rules with function words they found that this improves the results significantly by capturing more syntactic structure. These kinds of syntactic features perform significantly better than lexical features alone, giving the best performance on the ICLE (v.2) dataset at the time.

One key phenomenon observed by Wong and Dras (2011) was that there were distributional differences across parse production rules indicative of particular native languages. One example is the production rule `NP` → `NN NN` (*i.e.* a noun phrase can consist of two singular nouns, such as *country park*), which appears to be very common amongst Chinese speakers compared with other native language groups. The claim that the paper makes is that this is likely to reflect determiner-noun agreement errors, as that rule is used at the expense of one headed by a plural noun (`NP` → `NN NNS`, *e.g.* for *country parks*). Their intuition is that there might be coherent clusters of related features, with these clusters characterising typical errors or idiosyncrasies, that are predictive of a particular native language; in the SLA context, which we discuss in §2.4 and this could reflect the phenomenon of underuse or overuse.

Wong et al. (2011) investigated the use of Latent Dirichlet Analysis-based topic modelling to produce a feature set with lower dimensionality by clustering similar features. This approach, however,

[9] Contrastive Analysis is described later in §2.4.1.

did not outperform the other baseline systems as the authors reported that LDA-induced classification models perform worse than the full feature-based models. However, they did find some evidence of coherent feature clustering, which is an interesting result that bears further investigation.

Ahn (2011) also investigated the feasibility of NLI using machine learning techniques. The reported experiments were performed on a total of eight different native languages, with 200 texts per L1. Seven classes were from the ICLE corpus (Bulgarian, Chinese, Czech, French, Japanese, Russian and Spanish) and an eighth class of native English speaker writings was used from a corpus of native English essays called the LOCNESS corpus.[10] Using a Maximum Entropy classifier, they achieved their best result of 81.3% using character trigrams alone, although part-of-speech n-grams and syntactical features also worked relatively well. However, further investigation by the author revealed that the trigrams were in fact capturing topic words. When the topic-specific content words were removed from the feature set, the classification performance dropped, demonstrating that the previous results were artificially inflated due to topic bias in the corpus.

Kochmar (2011) was inspired by the error-based approach used in Wong and Dras (2009a). The author focuses on pairwise SVM classification on texts from the CLC FCE dataset which we earlier described in §2.3.1.3. The available native languages include five Germanic linguistic varieties (German, Swiss German, Dutch, Swedish and Danish) and five Romance languages (French, Italian, Catalan, Spanish and Portuguese), but the author focuses mainly on pairwise binary classification between language pairs. Similar to the previously reported results of Wong and Dras (2009a), the manually derived error-type features (*e.g.* spelling and determiners) that are annotated in the corpus did not yield any improvement over the baseline performance provided by the distributional features (character/word/part-of-speech n-grams and production rules).

Brooke and Hirst (2011) demonstrated that the most commonly used multiple-L1 learner corpus in NLI research at the time, the *International Corpus of Learner English*, was affected by topic bias. They did this by grouping the texts together according to topic to create two sets. One set was used for training and the other for testing and the roles were also reversed. This was compared against a random split of the same texts and showed that the topic-based split performed significantly worse ($p < .0001$) using several different core feature types. The authors hypothesize that this was due to the classifier not being able to take advantage of topic-L1 correlations. This bias was a confounding factor that was artificially inflating the results of experiments using this dataset. To address these issues, they discussed new methods of obtaining "cheap" corpora scraped from the web. The authors also noted that acquisition of such non-biased corpora would enable researchers to use lexical features, which had been previously avoided.

Motivated by their findings in Brooke and Hirst (2011), Brooke and Hirst (2012a) state that NLI "suffers from a relative paucity of useful training corpora, and standard within-corpus evaluation is often problematic due to topic bias." To address these issues they proposed the use of web-scraped L1 corpora and tested the utility of alternative sources of training data for NLI. The reported results of their study demonstrate that automatically translated word bigrams can achieve accuracies as high as 48.3% on a four-way classification task using the ICLE corpus. This result is much lower than previously reported studies that exclusively used within-corpus training and testing, but this is a tradeoff that ensures lower topic bias. Furthermore, their experiments also showed that increasing

[10] `https://www.uclouvain.be/en-cecl-locness.html`
`http://www.learnercorpusassociation.org/resources/tools/locness-corpus/`

the amount of training data resulted in a drastic improvement in accuracy (with boosts of up to 30%).

Torney et al. (2012) investigated the usefulness of psycholinguistic features for the task of NLI. These features had not been previously applied and the authors hypothesized that they would provide new information and insights as they are "based in psychoanalysis rather than computational linguistics". The features were based on the Linguistic Inquiry and Word Count (LIWC) tool which provides word counts in 80 different categories. Unlike other experiments, they used equal numbers of texts from all 16 languages in the ICLE corpus, with 241 texts per L1 class and a random baseline of 6.3%. Compared against other standard feature types (function words, character and POS n-grams), the LIWC feature set was found to be the best single feature type, achieving an accuracy of 50.3%. This accuracy was further increased when other standard features (such as character and part-of-speech n-grams) were added, which the authors claim is evidence that the LIWC features are sufficiently different from the others to improve classification accuracy. However, it must be noted that the LIWC dictionary[11] contains many topical content words,[12] making this experimental setup highly vulnerable to topic bias, particularly as it uses the ICLE corpus. This is substantial shortcoming to consider, and without additional experiments, it is difficult to assess the validity of these claims. This experiment is also notable for having used a very large number of L1s.

Wong et al. (2012) proposed the use of adaptor grammars in order to extend the character and part-of-speech n-gram features. In their experiment, adaptor grammars are used to discover n-grams of arbitrary length consisting of mixtures of part-of-speech tags and function words (*e.g.* `NN of NN`). The authors achieved the best n-gram results to date and showed that this composite feature can produce good NLI features.

Swanson and Charniak (2012) explored the use of automatically induced Tree Substitution Grammar fragments as classification features for NLI, further extending the work of Wong and Dras (2011). They compared two state-of-the-art methods for Tree Substitution Grammar induction and showed that features derived from both methods outperformed current state-of-the-art results. In comparatively evaluating the Bayesian and DoubleDOP induction methods on the ICLEv2 data, they found that the Bayesian approach resulted in the highest accuracy (78.4%).

Tetreault et al. (2012) proposed the use of classifier ensembles for NLI and performed a comprehensive evaluation of the feature types used until that point. In their study they used an ensemble of logistic regression learners using a wide range of features that included character and word n-grams, function words, parts of speech, spelling errors and writing quality markers. With regard to syntactic features, they also investigated the use of Tree Substitution Grammars and dependency features extracted using the Stanford parser. Furthermore, they also proposed using language models for this task and in their system used language model perplexity scores based on lexical 5-grams from each language in the corpus. The set of features used here was the largest of any NLI study to date. With this system, the authors reported state of the art accuracies of 90.1% and 80.9% on the ICLE and TOEFL11 corpora, respectively. Tetreault et al. (2012) also conducted cross-corpus evaluation, using the 7 common L1 classes between the ICLE and TOEFL11 corpora. Training on the ICLE data, they report an accuracy of 26.6%.

Bykh and Meurers (2012) explored additional use of lexical features and proposed the systematic use of all "recurring n-grams", an approach Meurers originally developed for corpus annotation error

[11] `http://www.liwc.net/descriptiontable1.php`

[12] *e.g.* the words *party*, *shopping* and *journey* are in the dictionary.

detection (Dickinson and Meurers, 2003, 2005), as features in NLI. They define recurring n-grams as all n-grams of any length that appear in at least 2 texts in the training data. They create recurring n-grams based on three feature types: words, parts of speech and Open-Class Part of Speech (OCPOS) which is a special case where only open-class words (nouns, verbs, adjectives and cardinal numbers) are converted to POS tags and closed-class words are left in their lexicalized form. For this study they used an SVM classifier on seven L1s from the ICLE. They report that their word-based n-grams achieved the highest accuracy (89.71%) followed by the OCPOS-based variety (80%). Interestingly, the authors claim that they could not find any evidence of topic bias in the ICLE, something which had been discussed by several other studies and could have potentially affected the accuracy of the word n-grams. We also note that this study was based on only a small amount of data (100 texts per L1, with 75 for training and 25 for testing).

Following these studies, it has since become well-established that syntactic features provide a strong discriminative cue about the writer's native language. The publication of these studies from 2009–2012 brought an unprecedented level of research focus and momentum to the area, culminating in the NLI Shared Task 2013.

2.3.3.3 The 2013 NLI Shared Task

The very first shared task focusing on Native Language Identification was held in 2013, bringing further focus, interest and attention to the field. The NLI Shared Task 2013[13] was co-located with the eighth instalment of the Building Educational Applications Workshop at NAACL/HLT 2013.

Two key challenges that have hindered the progress of this field are paucity of suitable data and the lack of a common evaluation framework.

As previously noted, most previous NLI research has used different L1s and differing numbers of classes and texts, along with variation in use of cross-validation and different data splits. The use of differing pre-processing techniques has also been a hindrance for cross-study evaluation of results. In absence of a common evaluation standard, it had been difficult for researchers to compare and contrast the various methodologies that had been explored. In this respect, the shared task also aimed to facilitate the comparison of results by providing a large NLI-specific dataset and evaluation procedure to all participants, thus enabling the direct comparison of results achieved through different means and methodologies.

The competition resulted in 29 entries from teams across the globe, 24 of which also submitted a reference paper describing their systems. The shared task featured 3 separate tracks or sub-tasks:

- **Closed-training**: The first and main task was the 11-way classification task using only the TOEFL11-TRAIN and optionally TOEFL11-DEV for training.
- **Open-training-1**: The second task allowed the use of any training data *excluding* TOEFL11-TRAIN and TOEFL11-DEV.
- **Open-training-2**: The third task reflects a more real-world scenario, allowing the combination of TOEFL11-TRAIN and TOEFL11-DEV with any other additional data.

Each team was allowed to submit up to 5 different systems for each task, allowing them to experiment with different feature and parameter variations of their system. Entries in all tasks were evaluated on the TOEFL11-TEST set.

[13] https://sites.google.com/site/nlisharedtask2013/home

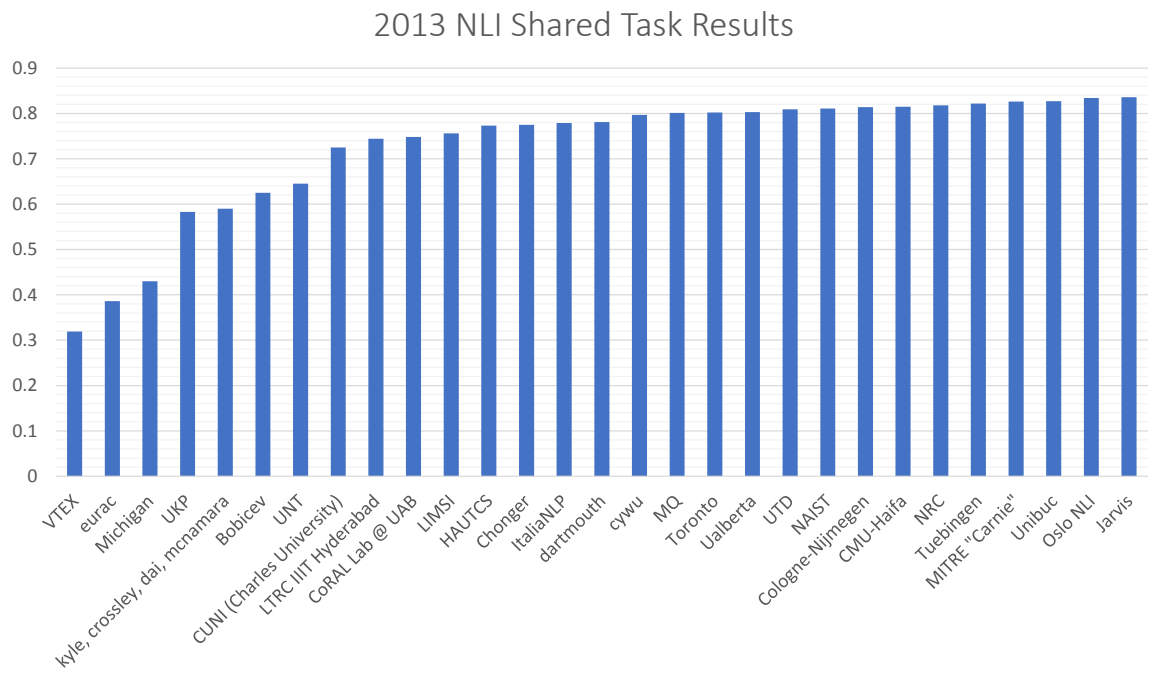

Figure 2.12: NLI 2013 shared task results showing the accuracy obtained by each team's best system.

Most teams only participated in the closed training task which attracted 29 teams competing and 116 submissions in total. Performances for each team's best system were in the range of 31.9–83.5%, as shown in Figure 2.12. There was no statistically significant difference between the results for the top 5 teams.

The two open training sub-tasks drew far less interest, with only four teams submitting systems. Accordingly, we do not include details relating to the open tasks in this section. The next section will review the systems that submitted a description paper. The aim is to identify features, learners and techniques that are commonly used. Our own entry will be described in Chapter 3.

2.3.3.4 Ensemble-based Shared Task Entries

A notable trend among the entries was the use of ensemble-based systems, which have been shown to achieve better results over systems based on single models. In this section we briefly review the systems that took this approach.

Gyawali et al. (2013) utilized lexical and syntactic features based on n-grams of characters, words and part-of-speech tags (using both the Penn TreeBank and Universal Parts Of Speech tagsets), along with perplexity values of character n-grams to build four different models. These models were combined using a voting-based ensemble of SVM classifiers. Features values were weighted using the TF-IDF scheme. In particular, the authors set out to investigate whether a more coarse grained POS tagset would be useful for NLI. They explore the use of the Universal POS tagset which has 12 POS categories in the NLI shared task and compare the results with the fine-grained Penn TreeBank (PTB) tagset that includes 36 POS categories. The highest accuracy of their system in the shared task is 74.8%, achieved by combining all features into an ensemble. The authors found that the use of coarse grained Universal POS tags as features generalizes the syntactic information and reduces the discriminative power of the feature that comes from the fine granularity of the n-grams. For example,

the PTB tagset distinguishes verbs into six distinct categories while the Universal POS tagset only has a single category for that grammatical class.

In the system designed by Cimino et al. (2013) the authors use a wide set of general purpose features that are designed to be portable across languages, domains and tasks. This set includes features that are lexical (sentence length, document length, type/token ration, character and word n-grams), morpho-syntactic (coarse and fine-grained part-of-speech tag n-grams) and syntactic (parse tree and dependency-based features). They report that they found distributional differences across the L1s for many of these features, including average word and sentence lengths. However, we note that many of these differences are not of a large magnitude, and the authors did not run any statistical tests to measure the significance levels of these differences. Using this feature set, they experiment with a single-classifier system as well as classifier ensembles, using SVM and Maximum Entropy classifiers. In their ensemble, they experiment using a majority voting system as well as a meta-classifier approach. The authors report that the ensemble methods outperform all single-classifier systems (by around 2%), and their best performance of 77.9% is provided by the meta-classifier system which used linear SVM and MaxEnt as the component classifiers and combined the results using a polynomial kernel SVM classifier. While the set of features used in this experiment is not widely different to other reported NLI research, their use of a meta-classifier is an interesting approach that warrants further study.

In their system, Goutte et al. (2013) used character, word and part-of-speech n-grams along with syntactic dependencies. They used an ensemble of SVM classifiers trained on each feature space, using a majority vote combiner method. To represent the feature values, they use two value normalization methods based on TF-IDF and cosine normalization. Their best entry achieved an accuracy of 81.8%, higher than many systems using the same standard features and more, demonstrating the effectiveness of using ensemble classifiers and appropriate feature value representation. The authors, like many others, also note that lexical features provided the best performance for a single feature in their system, but that this can be boosted by combining multiple predictors.

The MITRE system (Henderson et al., 2013) is another highly lexicalized system where the primary features used are word, part-of-speech and character n-grams. In this system, these features are used by independent classifiers (logistic regression, Winnow2 and language models) whose output is then combined into a final prediction using a Naïve Bayes model. Their best performing ensemble was 82.6% accurate in the shared task and the authors emphasize the value of ensemble methods that combine independent systems. Furthermore, the authors also optimized the parameters of their Naive Bayes model using a grid search over the development data.

Hladka et al. (2013) developed an ensemble classifier system using some standard features (lemma, word and part-of-speech n-grams, word skipgrams) with SVM classifiers. They obtained an accuracy of 72.5% in the shared task. They found that their ensemble, which is based on majority voting, outperformed other methods of combining the features. This is yet another piece of evidence pointing to the utility of using ensemble systems for NLI.

Another system that utilizes an ensemble is that of Bykh et al. (2013), where they used a probability-based ensemble. They use a set of 16 features, including recurring word-based n-grams, recurring OCPOS-based n-grams (as described in §2.3.3.2), dependencies, trees and lemmas. To combine the different feature types, they explored combining all feature into a single vector and also ensembles of SVM classifiers (each trained on a single feature type). Their best shared task perfor-

mance of 82.2% was achieved using an ensemble with all of their features. Their analysis shows that recurring word-based n-grams are the best performing single feature type, once again demonstrating the relevance and significant of lexical features in NLI.

2.3.3.5 Other Shared Task Entries

Abu-Jbara et al. (2013) used an SVM classifier trained on lexical and syntactic features. In addition to the standard features, they also included the number of unique word stems, punctuation and capitalization patterns, tense and aspect frequencies, and the usage of downtoners (a degree adverb such as *somewhat*) and intensifiers. Their best entry in the shared task achieved an accuracy rate of 43.3%. The authors note that the best features in their system were character, word and part-of-speech n-grams.

Daudaravicius (2013) took an approach that did not require any deep language processing, in an attempt to engineer features that would be easier to generalize to other languages. The author uses word length and suffix information and combines them into a new feature referred to as $n$end transformation, where the word length and N characters from the end are combined to create a new token. For example, the **3end** transformation for the word *make* is *4ake*. Next, n-grams of these $n$end features are extracted. As opposed to the common classification approaches taken by most other participants, the author chose instead to utilize a k-NN classifier, using the distance between the test and training documents to assign the language label. The best submission had an accuracy of 31.9% using this approach. This approach, while not achieving a competitive result in the shared task, is notable for being language-independent requiring very little information processing. We also note that using the word length and suffix information as separate features is also another way of using the same features without an additional transformation to fuse them together.

The system developed by Bobicev (2013) is also designed to be language-independent and require little linguistic processing of the data. This system is based on Prediction by Partial Matching (PPM), which is an adaptive finite-context method for text compression that uses a back-off smoothing technique for finite-order Markov models. It uses the last few observed symbols to predict the upcoming symbol, where symbols may either be characters or words. This method uses the input text in its original format and requires no feature engineering or extraction. The best entry for this system obtained an accuracy of 62.5% using a word-based model that significantly outperforms its character-based counterpart. While this compression-based classification approach is radically different to those adopted by most other entries, it should be noted that the results are not markedly different to those reported by others using word unigrams.

The winning entry for the shared task was that of Jarvis et al. (2013), with an accuracy of 83.6%. The features used in this system are n-grams of words, parts-of-speech as well as lemmas. In addition to normalizing each text to unit length, the authors also applied a log-entropy weighting schema to the normalized values, which clearly improved the accuracy of the model. An L2-regularized SVM classifier was used to create a single-model system. Furthermore, the authors employed their own procedure for optimizing the cost parameter (C) of the SVM. While they did not use a great number of features or introduce any new features for this task, we posit that their use of weighting schema and hyperparameter optimization gave their system an edge over their competitors, the majority of whom did not employ these techniques.

Li (2013) also used a linear SVM classifier with character and word n-grams, style features from authorship attribution (such as sentence and paragraph lengths) and the proficiency level information provided in the corpus as metadata. This system achieved an accuracy of 77.3%. The author notes that the style and proficiency features did not provide any improvements in their results.

The Nara Institute of Science and Technology (NAIST) Native Language Identification system by Mizumoto et al. (2013) uses an SVM classifier with a particular focus on feature selection methods for improving accuracy. Instead of using standard frequency-based feature weighting methods such as TF-IDF, they propose using a measure of **Native Language Frequency** (NLF) which is the number of languages (classes) in which a feature appears. The NLF value for this experiment ranges between 1–11, where 11 means a feature appears in at least one text in every language. Applying this to pre-existing features from the literature (character/word/part-of-speech n-grams, dependencies and TSG fragments), the system's best result in the shared task was 81.7%. This was achieved by using their proposed feature selection method by excluding features that appeared in just one or all languages ($1 < \text{NLF} < 11$) and binary feature representations.

Nicolai et al. (2013) also opt to use an SVM classifier with standard features, but also focus on developing new features based on cognate interference and spelling errors. They propose a new feature based on interference from cognates, words with a common linguistic origin from an ancestor language, positing that interference may cause a person to use a cognate from their native language or misspell a cognate under the influence of the L1 version. For each misspelled English word, the most probable intended word is determined using spell-checking software. The translations of this word are then looked up in bilingual English-L1 dictionaries for several of the L1 languages. If the spelling any of these translations is sufficiently similar to the English version (as determined by the edit distance and a threshold value), then the word is considered to be a cognate from the language with the smallest edit distance. The authors state that although only applying to four of the 11 languages (French, Spanish, German, and Italian), the cognate interference feature improves performance by about 4%. Their best result on the test was 81.73%. While limited by the availability of dictionary resources for the target languages, this is a novel feature with potential for further use in NLI. An important issue to consider is that the authors' current approach is only applicable to languages that use the same script as the target L2, which is Latin and English in this case, and cannot be expanded to other scripts such as Arabic or Korean. The use of phonetic dictionaries may be one potential solution to this obstacle.

Wu et al. (2013) chose to explore the role of the feature representation in their system, using only word unigram and bigram features to comparatively evaluate the performance of binary and frequency-based feature value representations. Using an SVM classifier, their system achieves an accuracy of 79.7% and the authors find that binary representations provide better performance in their system. This finding is consistent with previously reported results (Brooke and Hirst, 2012b). They also report that the inclusion of punctuation in their feature set was useful, but that the proficiency information was not predictive.

The system developed by Brooke and Hirst (2013) also employed a single SVM classifier. In addition to the standard features used in their previous work (function words, character/word/part-of-speech n-grams, dependencies, context-free productions, and "mixed" POS/function n-grams), they also introduce several new features such as POS-abstracted n-grams, TSG fragments and psychological word category association information from psycholinguistics. However, their testing demonstrates

that none of these yield any improvement in classification accuracy. They experimented with using frequency cutoff thresholds as a feature selection method, but found that optimal performance was achieved with no feature selection at all. They also experiment with proficiency-segregated models, but like several other authors, they did not find them to be helpful in improving classification.

With a strong focus on syntactic features, Swanson (2013) contrasts the performance of NLI using five different representations of the syntactic parsing results. Tree Substitution Grammars are used to parse the training texts and generate a set of TSG rules which are used as binary features for classification. Swanson evaluates the role of the syntactic paradigm by comparing the performance of five different variations of the output of the parsing process. The five formalisms used are plain Berkeley Parses, Berkeley Parses with split symbols, dependency parses, dependency parses without arc labels, and the heuristic annotations which are internal to the Stanford Parser. Tree Substitution Grammars for the shared task training data were induced using each of these five output formats and classification accuracy was compared using a MaxEnt classifier. While the accuracies for the individual models ranged between 69–74%, the system's best accuracy of 77.5% was achieved by averaging the results of all models together. This result is also noteworthy for being one of the highest results achieved using only a syntactic feature. The authors conclude by saying that instead of attempting to find the optimal formalism for the task, NLI can benefit from using a range of syntactic representations, further noting that there exist several other syntactic forms that have not been explored in their work.

Another system that focuses on exploring feature weighting schemas is the entry by Gebre et al. (2013) which uses standard NLI features (character/word/part-of-speech n-grams and spelling errors). The raw frequency values for these features are first weighted using the well-known TF-IDF scheme and then normalized (using the l^2 norm) to account for differing text lengths. Using this weighting technique and a few common features, their entry was able to achieve a competitive accuracy of 81.40%, which the authors attribute to their use of the weighting scheme.

Lynum (2013) created a system trained with a large number of lexical features. The author was motivated by the possibility that the efficient application of simple lexical features would overcome the need for an NLI system to learn the syntactic and morphological differences in the non-native texts. In this spirit, only features derived from the surface forms of the available texts were used. Features include word unigrams and bigrams, character n-grams and suffix bigrams based on the last four characters of each word. Furthermore, the feature values were weighted using the TF-IDF weighting scheme. The learner used in this experiment was a linear SVM and the SVM model hyperparameters were also optimized over 5-fold cross-validation on the training data. This system achieved an accuracy of 83.4% in the shared task, a strong result given the limited features types used and the parsimonious nature of the approach.

Lavergne et al. (2013) designed their NLI system based on a Maximum Entropy classifier, training it with a set of basic (character and part-of-speech n-grams) and complex features (spelling mistakes, grammatical mistakes and lexical preference). The best result for their shared task entry was an accuracy of 75.6% on the test data. Based on their experiments the authors note that while their best results come from using the complete set of features, the basic n-gram features are the best performing and no other type of feature provides significant improvement when added to the system. The authors also observe that two specific language pairs have high rates of confusion: Hindi-Telugu and Japanese-Korean. In an attempt to improve the classification accuracy of their models, they

propose the use of a two-step classifier targeted at these specific hard to distinguish pairs. To this end they created additional classifiers for each and documents classified as one of the languages in the difficult pairs were re-processed by the second classifier. However, this approach only results in a 0.17% improvement in accuracy.

Lahiri and Mihalcea (2013) experimented with a range of classifiers (SVM, Naive Bayes, 1-nearest-neighbor (1NN), J48 decision tree, and AdaBoost) utilizing a set of baseline n-gram features (words, characters and parts-of-speech) as well as a novel set of features derived from Word Networks. Word Networks are graphs of unique words found in a text where each vertex is a word and edges connect words based on the graph building criteria. In this work, the authors created unweighted directed edges between words that appeared consecutively. Once the network has been constructed, a set of graph-theoretic features based on neighbourhood size, coreness and degree are extracted. While this is a novel feature for NLI, it was unable to outperform the baseline feature set. Their best result of 64.5% was achieved using the SVM classifier and baseline features, while the word network features performed at 63%. They note that word n-grams performed the best, followed by POS and character n-grams. Further experiments are needed to determine whether the word network features are complementary to the baseline n-gram features and if combining them all together could yield even better results, particularly by using classifier ensembles.

Kyle et al. (2013) explored the efficacy of an approach based on categories of key n-grams. They created several such categories that included grammatical, rhetorical, semantic and syntactic key n-grams. The authors had previously assessed the efficacy of these features within an automatic writing evaluation system where they split essay texts into introduction, body, and conclusion paragraphs, and further separated these into high and low proficiency categories based on the essay's overall score. Next, they used "keyness analysis" to identify n-grams that occurred significantly more often (positive keyness values) in paragraphs of a specific type (*e.g.* in the introduction) from high scoring essays than the same type of paragraphs from low-scoring essays. The positively and negatively key n-grams for each paragraph type were then further separated into categories based on their rhetorical, syntactic, grammatical, and cohesive characteristics and then used as variables in a multiple regression to generate a model accounting for 24%–33% of the variance in essay scores. Motivated by their previous success, they aimed to assess the performance of such a system on a prompt and proficiency-controlled corpus and examine if this model could be applied to predict the author's native language. Key n-grams were identified on a per-language basis, by comparing the texts against a reference corpus comprised of the texts of the other 11 languages. The n-grams were then manually categorized into lists by two trained linguists with experience in the area of second language writing. Using discriminant function analysis for classification, their highest accuracy in the shared task was 59.0%. Analysing their results, the authors observe a number of findings that are potentially attributable to language transfer. Given that this methodology is based on a manual categorization of the word n-grams used by other systems, the authors also note that it is still unclear whether this approach is more or less effective and further research will be required to determine whether the costs for the manual creation of categorical n-gram lists is warranted.

Popescu and Ionescu (2013) approached the shared task using machine learning methods that work at the character level, using string kernels and a kernel based on Local Rank Distance (LRD). This is a machine learning oriented methodology that treats the text as a sequence of symbols, thus making it another language-independent approach which is not based on any specific linguistic theory. The

authors' best entry ranked third in the shared task with an accuracy of 82.7%. This was achieved by combining the two kernel methods in a classifier using Kernel Ridge Regression (KRR), which the authors found to outperform an SVM classifier. String kernels have previously been shown to be highly useful for text categorization (Lodhi et al., 2002) and the authors, who have successfully applied these methods to authorship identification problems in the past, believe that these good results are due to the fact that a string kernel of length 5 creates a high dimensional feature space that embeds all substrings of length 5. In this sense, the features captured by the string kernel are quite similar to character 5-grams, with the exception that the string kernel is not aware of and does not use word boundary information, as is the case with the common method of extracting character n-grams that does not cross word boundaries. Consequently, the string kernel is implicitly taking into account function words, short content words, character 5-grams, word stems and morphological information such as suffixes. The kernel does not play a role in selecting the most appropriate feature, but leaves it to the learning algorithm to select the most discriminative ones from within the very large feature space. This is a strong result given that unlike most other entries their system disregards linguistic information about words, phrases, semantics, syntax and grammar.

Tsvetkov et al. (2013) used standard text classification techniques based on multiclass logistic regression, combining individually weak predictors to classify the most probable native language of the texts. Using a feature set that draws heavily on prior work in general text classification and authorship identification their arsenal of feature types includes, amongst others, word/character/part-of-speech n-grams, function words, and Brown clusters.[14] This system obtained an accuracy of 81.5% in the shared task, using a single model with all of these features. Their use of Brown clusters is notable as the first application of the technique for NLI. The Brown word clustering algorithm can be used to generate hierarchical clusters of similar words, and has previously been used as a feature in various NLP tasks. For use in NLI, the authors report an accuracy of 72.26% using counts of Brown cluster unigrams and bigrams in each document. This is an interesting result, although further clarification is required to determine whether the information provided by this feature is redundant or orthogonal in relation to other lexical features such as word n-grams.

2.3.3.6 Shared Task Summary

As a whole, the first NLI shared task can be considered a success, drawing a large number of entrants and experts from not only Computational Linguistics, but also SLA. In reviewing the shared task results and entries, we note a number of trends.

First, we observe that most systems had many commonalities. The majority of entrants used a very similar set of standard features and machine learning methods. Similar to much of the previous literature, the most commonly used features were character, word, and part-of-speech n-gram features (usually with $n <= 4$) and syntactic features. Support Vector Machines were also the most commonly used machine learner. While many teams drew heavily on prior work, others were notable for their use of novel and unique approaches, such as the use of string kernels by Popescu and Ionescu (2013).

With so many similar entries, we need to look at the more nuanced differences between systems to better understand the components that enabled some teams to achieve better results.

[14] Brown clustering (Brown et al., 1992) is an unsupervised method for clustering words into classes based on the contexts in which they are observed. Words are hierarchically grouped into a binary tree of classes. It has been shown that this clustering method produces effective features for discriminative models.

With regards to the machine learning techniques, Support Vector Machines were generally the most successful. Another step that was taken by most of the top entries (particularly the top 5), but not by most of the other entries is that of hyperparameter tuning and optimization. The most accurate systems used the training and development data to optimize the parameters of their learning algorithm and used the parameters for classifying the test data. We also note that the highest performing entries also made use of classifier ensembles, instead of using a single learner.

Secondly, we note that the most accurate systems employed some form of weighting schema for the feature values in addition to normalizing them to unit length. This was the case for several of the teams in the top 10 and those who applied this technique seemed to have obtained better results then those who did not, but had comparable systems and feature sets. We believe that this is a potentially useful component that holds promise for boosting the performance of current NLI systems and features. Additional research that directly assesses the role of feature weighting schemas is needed to determine their utility.

Several entries attempted to use the proficiency-level meta-data included with the corpus to aid the classification, for example the systems by Kyle et al. (2013), Brooke and Hirst (2013) and Malmasi et al. (2013). However, none of them were able to successfully leverage this information to improve their system accuracy. One potential issue here is that while the TOEFL11 corpus is balanced for topic and language, it is not balanced for proficiency levels. Overall, the three proficiency levels (High, Medium and Low) are not equally represented, with over 85% of the texts falling in the High and Medium categories (though not equally). This imbalance may be what makes it hard to establish and exploit any proficiency-related patterns within this specific corpus. However, experiments with a proficiency-balanced corpus may yield different results.

Several authors have confirmed that stylistic features (such as the number of characters/words/sentences/paragraphs or the average word/sentence length) are not useful for the task of Native Language Identification (Nicolai et al., 2013; Li, 2013).

2.3.3.7 Recent Work

Swanson and Charniak (2013) develop "a method for effective extraction of linguistic patterns that are differentially expressed based on the native language of the author". To this end they examine both relevancy and redundancy of NLI features through a number of ranking metrics (including the chi-square statistic). They then extend a Bayesian induction model for TSG inference based on a supervised mixture of hierarchical grammars, in order to extract a filtered set of more linguistically-informed features that could benefit both NLI and SLA research; an aim of their research was to find relatively rare features that are nevertheless useful for L1 prediction.

Perkins (2014) presents an NLI study of L1 Persian speakers of English using data obtained from publicly accessible blogs instead of learner compositions. Instead of using the supervised learning approach, this work is based on statistical analysis of manually coded linguistic features which include part-of-speech categories along with additional flags that encode marked choices, ordering, positioning, etc. The first study compared the English writings of L1 Persian authors against natives, finding a statistically significant difference in the linguistic features used by the two groups. The second study also showed that such differences also exist when comparing against L1 Azeri and L1 Pashto authors. Motivated by past forensic cases that involved perpetrators making an effort to disguise their true L1, the third study focuses on the interesting problem of comparing L1 Persian speakers against a group

impersonating Persian speakers. This involved eliciting data by asking participants to disguise their linguistic background. Results indicate that there were considerable differences between the true and disguise groups.

Bykh and Meurers (2014) explore the use of lexicalized and non-lexicalized phrase structure rules for NLI. They show that the inclusion of lexicalized production rules (*i.e.* preterminal nodes and terminals) provides improved results. In addition to the standard normalized frequency and binary feature representations they also propose two new representations based on a "variationist sociolinguistic" perspective. Although they show that these representations outperform the normalized frequency approach, they do not compare this to other representations which have been shown to improve NLI accuracy, such as TF-IDF. They combine their lexicalized production rules feature with additional surface n-gram features in a tuned and optimized ensemble, reporting an accuracy of 84.82% on the TOEFL11-TEST set.

Ionescu et al. (2014) extend the previous work of Popescu and Ionescu (2013) which used string kernels to perform NLI using only character n-gram features. One improvement here is that several string kernels are combined through multiple kernel learning. The authors also perform parameter tuning to select the optimal settings for their system. They report an accuracy of 85.3% on the TOEFL11-TEST set, 1.7% higher than the winning shared task system. It is also suggested that using higher order n-grams, with n as high as 8, improves classification accuracy.

Nicolai and Kondrak (2014) investigate the issue of how and if L1 phonology is manifested in L2 texts. They investigate the source of L1 differences in the relative frequencies of character bigrams since this is the feature that is most commonly assumed to capture phonological transfer, as claimed by Tsur and Rappoport (2007) which we discussed in §2.3.3.1. They propose an algorithm to identify the most discriminative words and subsequently, the extracted bigrams that correspond to these words. Empirical evidence from their experiments suggests a different explanation to the inter-L1 differences in character bigram distributions. They found that removing a small set of highly discriminative words greatly degrades the accuracy of a bigram-based classifier. Based on this they conclude that bigrams capture differences in word usage and lexical transfer rather than L1 phonology.

2.4 Second Language Acquisition: Cross-linguistic Influence

In this section we look at how a learner's first language can influence the acquisition of other languages. It is important to describe these theories as what most of the NLI systems described in the previous sections are attempting to detect is this type of L1 influence. We will also return to this link between NLI and L1 influence in Chapter 6, examining it in a more detailed manner.

Cross-linguistic influence (CLI), also referred to as language transfer, is one of the major topics in the field of Second Language Acquisition (SLA). It has been said that being a speaker of some specific native language (L1) can have direct and indirect consequences on an individual's usage of some later learned language (Jarvis and Crossley, 2012, p. 1), and this is the effect that is studied in CLI.

However, linguists have historically been more focused on investigating a single language, and often on some particular aspect such as phonology, syntax or morphology. Few have entirely focused on making cross-linguistic comparisons or the influence of languages on each other in multilingual individuals (Myers-Scotton, 2005, p. 11).

Studies of bilingualism have led to two extreme positions (Myers-Scotton, 2005, p. 12). One view is that considering multiple languages would complicate matters and thus obscure our view of human language. The second position posits that studying bilinguals presents us with a opportunity to better understand the structures of these languages by observing how they interact with and influence each other. In this sense, the study of CLI through analysing non-native text can be helpful in informing second-language pedagogy

Of the proposed models of cross-linguistic influence, two of the earliest and most widely recognized ones are *Contrastive Analysis* and *Error Analysis*.

2.4.1 Contrastive Analysis and Error Analysis

Lado (1957) proposed *Contrastive Analysis* as one of the first methods to study language transfer effects in learners. This methodology was focused on comparing the native and non-native languages to identify to those differences that could give rise to learning difficulties.

Based on this work, Lado advanced the Contrastive Analysis Hypothesis which posits that those elements which are similar to the native language will be easy to learn, and those elements that are different will be difficult. This hypothesis predicts that learners will commit more errors related to those linguistic aspects that differ the greatest. It also predicts that the "interference" (or negative transfer) from the native language would increase with this degree of dissimilarity. The term interference can refer to a number of phenomena that can occur when languages come into contact. These effects may include lexical and syntactic transfer, amongst other things.

This approach gained significant traction over the next decade, but through continued research it became evident that some errors were being committed by learners from varying linguistic backgrounds in way that were not predicted by the model. Furthermore, it was also noted that some errors which were being predicted by this model were not actually observed. Another issue with this approach was that performing a comprehensive comparison of a language pair was a huge undertaking that required significant resources.

These shortcomings led Corder (1967) to propose *Error Analysis* as an alternative methodology that is centred on the typology of learner errors. This model extends the Contrastive Analysis approach by proposing that there are several types of errors which also include the "interference" errors predicted by Lado's framework. This approach shifted the research focus to the actual errors being produced by learners and "provided a methodology for investigating learner language" (Ellis, 1994, 48).

Error analysis is based on the premise that errors are a reflection of a learner's current knowledge or competence of a language at their current stage of the language acquisition process. It postulates that these errors fall into several distinct categories, some of which are universal and not necessarily due to interference from the mother tongue (Richards, 1971). Such universal errors may be due to interference of the various patterns and irregularities with the target language and thus considered to be developmental in nature.

2.4.1.1 Error Types

These systematic errors were classified into three categories by Richards (1971): developmental errors, intralingual errors and interference (interlanguage) errors.

Developmental Errors These errors are similar in nature to those observed during the acquisition of the first language. Learners generally have limited knowledge about the workings of a new language but attempt to leverage this narrow knowledge to form hypotheses about the new language. As a result of their limited comprehension and knowledge, these hypotheses are often faulty and their application results in developmental errors (Richards, 1971). Their prevalence decreases with time as a learner's hypotheses and learned rules are refined through practice and experience.

Intralingual Errors Richards (1971) argues that one type of common error is caused by the learning strategies that learners use in their study of a new language. The manifestation of these errors can be part of the rule learning process, caused by the incorrect or incomplete application of the rules. For example, overgeneralization can occur when the application of a previously learned rule in an inappropriate context yields an incorrect linguistic structure. Such errors commonly result in noun-verb and tense agreement errors.

Interference (interlanguage) Errors These types of errors can be attributed to the role of the native language. They can generally be predicted by examining and contrasting the linguistic subsystems (such as the syntactic or morphological systems) of both languages. For example, in examining a language pair such as Russian-English we may predict that Russian learners of English may have difficulties in correctly using determiners due to the absence of these rules in their native language.

The scope of such studies may also be narrow and more focused on particular facets of language use. For example, Granger and Tyson (1996) investigated how English learners from various backgrounds used English connectors. Looking at texts written by Chinese, German, Dutch, and French native speakers, they found evidence of overuse and underuse of individual connectors. Yang and Huang (2004) examined the impact of the absence of grammatical tense in Chinese on the acquisition of the English tense-aspect system for native Chinese speakers. They found that the Chinese system of expressing temporal information may reinforce the learners' initial tendencies of using pragmatic and lexical devices to indicate tense, which would be a type of negative transfer. Even with early instructions on grammatical tense, they found a very slow shift towards the use of grammatical devices due to the L1 interference which can be quite strong due to the magnitude of the difference between the two languages' tense systems.

These transfer errors can also be observed as part of the orthographic system. In one study on such transfers, Sun-Alperin and Wang (2008) reported that native Spanish speakers committed more English spelling errors which were related to the English vowel system. Intuitively this finding is consistent with the hypothesis that could be drawn by contrasting the orthographic and phonological systems of the two languages. While Spanish is consistent and highly regular in its spelling and associated pronunciation, the link between English orthography and phonetics is not highly consistent, particularly with diphthongs and triphthongs. This difference may lead to negative transfer where learners may produce incorrect orthographic forms of English words which more closely resemble the Spanish spelling of the word's phonetic representation.

2.4.2 Transfer Effects

Having discussed how CLI has been conceptualized by researchers, we now turn our attention to a discussion of how these transfer effects manifest themselves in the language production of a learner. These manifestations include positive transfer, overuse and avoidance.

Positive Transfer This type of transfer is generally facilitated by similarities between the native tongue and second languages. The transfer effect can also differ for the various subsystems of a language. The degree of similarity between two languages may vary in their vocabulary, orthography, phonology or syntax. For example, high similarities in one aspect such as vocabulary may facilitate high levels of transfer in language pairs such as Spanish-Portuguese or German-Dutch, but not as much in other facets (Ellis, 2008). Such effects can also be observed in orthographic systems where Chinese and Japanese native speakers may find it easier to learn each other's languages in comparison with those that speak a language which utilizes a phonetic alphabet.

Avoidance The underutilization of particular linguistic structures is known as avoidance. While the existence of this phenomena has been established, the source of this underproduction is a debated topic (Gass and Selinker, 2008).

One possible explanation is that avoidance is chiefly caused by the dissimilarities between two languages. Evidence for this hypothesis was provided from a seminal experiment by Schachter (1974) which demonstrated that English learners of Japanese and Chinese backgrounds made significantly fewer relative clause errors than their Farsi and Arabic speaking counterparts. This was not because Japanese and Chinese had syntactic structures more similar to English (in fact, the opposite is true), but rather because they were avoiding the use of such structures. Another reason for avoidance may be the inherent complexity of the structures themselves (Gass and Selinker, 2008).

Overuse The abovementioned avoidance or underuse of specific linguistic structures may result in the overuse of other structures. In learners, this may manifest itself as the reluctance to produce more complex constructions, instead opting to use combinations of simple sentences to express their ideas.

Ellis (2008) mentions that overuse can occur due to intralingual processes such as *overgeneralization*. This usually occurs when regular rules are applied to irregular forms of verbs or nouns, such as saying *runned* or *shoeses*.

2.4.3 Lexical Transfer

The transfer effects associated with lexical items have been previously noted by researchers and linguists (Odlin, 1989). These effects are mediated not only by similarities in word forms, but also word semantics and meanings. This can lead to a great deal of positive transfer between similar languages that expedites the learning process. For example, the closeness between the Spanish words *presidente/construcción/leyenda* and their English equivalents *president/construction/legend* will play a facilitatory role in the language acquisition process. This concept has been noted as far back as the work of Sweet (1899), who stated:

> Mastering the vocabulary of most European languages means simply learning to recognize a number of old friends under slight disguises, and making a certain effort to learn a

> residue of irrecognizable words, which, however, offer less difficulty than they otherwise would through being imbedded in a context of familiar words. The higher vocabulary of science, art, and abstract thought hardly requires to be learnt at all; for it so consists either of Latin and Greek terms common to most European languages or of translations of them.
>
> It is very different with a remote disconnected language such as Arabic or Chinese. The abstract vocabulary of Arabic shows Greek influence, although this affords very little practical help; but the terminology of Chinese philosophy and science is independent of Western influence, so that every extension of the vocabulary requires a special effort of memory and reasoning. The task of mastering such languages is literally an endless one. Enough Arabic grammar for reading purposes is soon acquired, the construction being always perfectly simple — at least in ordinary prose, but the student may read one class of texts for years, and then, when he proceeds to another branch of the literature, he may find that he can hardly understand a word, this being almost entirely the result of the unfamiliarity of the new vocabulary required.

This observation has been confirmed by studies which have found that lexical similarities aid learners with both language comprehension and production. One such study is that of Ard and Homburg (1983) who compared the ESL performances of Arabic and Spanish speaking students. The Spanish speaking students were far more successful in vocabulary questions, particularly for words that had very similar spellings in both languages. This provided empirical evidence for the advantageous effects of recognizing and using cognates, showing that it provided the Spanish speakers with an edge over the Arabic speakers. This ability to establish links between L1 and L2 lexical forms allows learners to develop a cognate strategy to aid with word production and recognition (De Groot and Poot, 1997; Meara, 1993).

We should also make it clear that this notion of lexical transfer is different from that of *loan words*. When speakers of two languages are in contact with each other they may learns some words and phrases from each other's languages. One group may also take some words and integrate them into their own language. Words from one language that appear in another are called *lexical borrowings*. This process usually occurs with nouns, where they are taken from a donor language and incorporated into a recipient language. The results of this process can be seen in the presence of Russian load words in the languages of former Soviet republics and other Eastern European countries. Many Chinese words have also been taken into South-eastern Asian languages. Such changes may be motivated by the fact that there is something more desirable or attractive about the other language, usually related to the prestige or other attributes of its speakers, including demographic, economic and socio-political considerations. It is because of these factors that English is the foremost source of borrowings into languages everywhere, particularly in the domains of science and technology.

These transfer effects can also be negative, often observed under the phenomenon of "false friends". Well known to many language teachers, this effect occurs when two languages with various cognates also have a subset of words that while morphologically similar, are not semantically related. Examples include the Spanish verb *realizar* (to carry out) which English L1 learners may confuse with *to realize* or the adjective *embarazada* (pregnant) which they often associate with *embarrassed*. In contrast with the facilitatory effects of true cognates, these can provide learners with difficulties. However, this is offset by the fact that in most languages the number of true cognates is much greater than that

of false ones (Hammer and Monod, 1976). A cognate dictionary of English-French was compiled by Hammer and Monod (1976) who reported that the ration of true to false cognates was roughly 11:1.

Researchers have also noted that the identification of true cognates is not always sufficient for ensuring their correct usage. There may be grammatical differences in how they are used within each language, and this can lead to other errors. A cognate verb present in two languages may employ a reflexive construction in one, but not the other, and not observing this grammatical restriction may lead to the erroneous application of the verb.

The degree of cognacy affects the cognate strategy employed by learners. Meara (1993) identified four such patterns, each with their own difficulties. The first pattern occurs when there are a large number of cognate-pairs, leading to considerable facilitation, for example in the case of Italian-Spanish. In the second case, the number of cognates may be few and such similarities may not be recognized or perceived by learners, resulting in very little facilitation. Some such language pairs are Turkish-Swedish or Arabic-Japanese. The third scenario is a mixed one where cognates are mostly present in the most common and frequent vocabulary, but rarely occur in the less frequent words. The final pattern occurs when the use of cognates is limited to specific domains and topics.

Another important condition for these effects is that learners must perceive and be aware of the presence of such cognates. Even when a considerable number of such words exist, not all learners may be aware of their utility. Researchers have noted that simply by training students to take note of cognates and use them helps them improve their vocabulary. In one such case Otwinowska-Kasztelanic (2009) used this method to raise awareness of cognate vocabulary as a strategy in teaching English to Polish adults.

It has been shown that this type of transfer is more likely to occur with nouns. Researchers posit that this is because the transfer of syntactic structures involves different cognitive processes.

2.4.4 New Directions in CLI Research

With the advancement of technologies in corpus linguistics and Natural Language Processing, researchers have been able to take new approaches to CLI research using large datasets and statistical models.

Jarvis (2012, pp. 5–10) groups methods for investigating CLI into two broad categories, the comparison-based and detection-based approaches, each with its own strengths and weaknesses.

The comparison-based approach In this approach, we generally focus on observing particular differences between two languages. For example, we may examine how two languages differ in their use of pronouns, articles or grammatical case. Once these differences have been identified, the next step is then to determine the extent to which, if any, these linguistic dissimilarities or variations impact L2 production. This process is carried out by examining writings of non-native speakers, with a particular focus on errors.

The detection-based approach This method takes the opposite approach in the sense that instead of beginning with a set of hypotheses, we instead canvas the available data to identify regularities and patterns that are distinct to writers from a particular L1 background. Once the significance of these patterns have been established through statistical testing, we can then hypothesize about the underlying linguistic causes of these intergroup differences. With respect to the field, Jarvis (2012)

then suggests that NLI is an innovative approach to tackling problems that SLA researchers have been looking at for a long time.

In essence, we can broadly consider these approaches to be based on deductive and inductive reasoning, respectively. The top-down, deductive approach entails the formulation of theories and the use of data to confirm them. Conversely, the bottom-up inductive approach uses observed patterns to form tentative hypothesis which are then tested for linguistic significance.

Within the field of linguistics, these methodologies are commonly referred to as corpus-based (deductive) or corpus-driven (inductive) approaches. A corpus-driven methodology is one that uses an inductive conceptualization in the analysis of specialized discourses (Groom, 2010, p. 60), made possible by the advent of advanced computational platforms, hardware, algorithms and digitized corpora. In this paradigm we begin with a neutral stance, free of any presuppositions, a priori knowledge or linguistic bias. Instead, we rely on information extracted from our corpus data (such as ranked frequency lists, n-grams, parse trees, keywords and concordances) to drive our analysis. Then, through inductive reasoning (bottom-up logic), we form linguistic hypotheses based on observed examples. This approach is contrasted by the "top-down" corpus-based methodology which uses corpora to test pre-existing linguistic theories and hypotheses (Tognini-Bonelli, 2001, p. 65). A detailed discussion of the differences between these two approaches can be found in Römer (2005, p. 8). In the corpus-driven approach, the text is the central component and the derived evidence is processed systematically. In exploring a language variety in this manner, we process this text to extract observations in which we can find patterns and commonalities.

Advances in computer hardware as well as the software that have enabled the development of large-scale digital corpora and Natural Language Processing methods have made the application of the corpus-driven approach much more feasible in the last two decades. Researchers have access to an increasingly large amount of electronic corpus resources and advanced software that facilitates the rapid analyses of texts using a steadily increasing arsenal of linguistic features.

2.5 Chapter Summary

In this chapter we presented a comprehensive survey of NLI as well as a brief discussion of wider areas of related work, including CLI. Continuing from here, we try to better understand NLI (its features, task structure, potential for improvement and application to practical tasks) in Chapters 3–6. Much of work and resources described in this chapter can be used to develop advanced models — *e.g.* via NLI — for extraction of evidence that can be attributed to CLI. In Chapter 6 we combine all of these elements and try out the detection-based direction, noted in §2.4.4, in an attempt to formulate language transfer hypotheses from large-scale learner data. We then assess whether the features obtained via this approach can help in a related task in Chapter 7.

Chapter 3

Native Language Identification

The work undertaken in this thesis is built upon a classification system for performing Native Language Identification. In this chapter we present our work in building such a system, which began with building our entry for the 2013 shared task. We used this as an opportunity to explore, *inter alia*, ensemble classifiers, different feature types and granularity of tagsets. In the second part of this chapter we explore feature *diversity*, the level of independence between the predictions provided by different feature types. We propose one approach to measuring this interaction between features and highlight several interesting trends that can help better understand the information being captured by the feature types.

Chapter Contents

The work presented in this chapter has been published as:

- Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. NLI Shared Task 2013: MQ Submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1716`
- Shervin Malmasi and Aoife Cahill. Measuring Feature Diversity in Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 49–55, Denver, Colorado, June 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-0606`

3.1 NLI Shared Task System

In this section we describe the system we built and experiments we conducted as part of the 2013 NLI Shared Task, which we detailed in §2.3.3.3.

Among the efflorescence of work on Native Language Identification (NLI) in the period leading up to the 2013 shared task, there were two trends in particular that we considered in building our submission. The first is the proposal and use of new features that might have relevance to NLI: for example, Wong and Dras (2011), motivated by the Contrastive Analysis Hypothesis (Lado, 1957) from the field of Second Language Acquisition, introduced syntactic structure as a feature; Swanson and Charniak (2012) introduced more complex Tree Substitution (TSG) structures, learned by Bayesian inference; and Bykh and Meurers (2012) used recurring n-grams, inspired by the variation n-gram approach to corpus error annotation detection (Dickinson and Meurers, 2003). Starting from the features introduced in these papers and others, then, other recent papers have compiled a comprehensive collection of features based on the earlier work — Tetreault et al. (2012) is an example, combining and analysing most of the features used in previous work. Given the relatively short timeframe of the shared task, there seemed to be not much mileage in trying new features that were likely to be more peripheral to the task.

A second trend, most apparent in 2012, was the examination of other corpora besides the International Corpus of Learner English used in earlier work, and in particular the use of cross-corpus evaluation (Brooke and Hirst, 2012b; Tetreault et al., 2012) to avoid topic bias in determining native language. Possible topic bias had been a reason for avoiding a full range of n-grams, in particular those containing content words (Koppel et al., 2009); the development of new corpora and the analysis of the effect of topic bias mitigated this. The consequent use of a full range of n-grams further reinforced the view that novel features were unlikely to be a major source of interesting results.

We therefore concentrated for the most part on two areas: the use of classifier ensembles, and the choice of part-of-speech tags. With classifier ensembles, Tetreault et al. (2012) noted that these were highly useful in their system; but while that paper had extensive feature descriptions, it did not discuss in detail the approach to its ensembles. We therefore decided to examine a range of possible ensemble architectures. With part-of-speech tags, most work has used the Penn Treebank tagset, including those based on syntactic structure. Kochmar (2011) on the other hand used the CLAWS2 tagset,[1] which is much richer and more oriented to linguistic analysis than the Penn Treebank one. Given the much larger size of the TOEFL11 corpus used for this shared task than the corpora used for much earlier work, data sparsity could be less of an issue, and the tagset a viable one for future work.

The description of our submission is therefore in several parts. We begin with a description of classifier ensembles in §3.1.1, followed by descriptions of ensemble combination methods in §3.1.2. In §3.1.3 we present the preliminary experiments we conducted to guide the development of our entries with a focus on the ensemble architectures we investigated, the features we used (which are for the most part those applied in much of the previous work) and other experiments we performed on the training data; in §3.1.4 we present our entries and their performance in the shared task; and in §3.1.5 we discuss some of the interesting characteristics of the data we noted during the shared task.

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

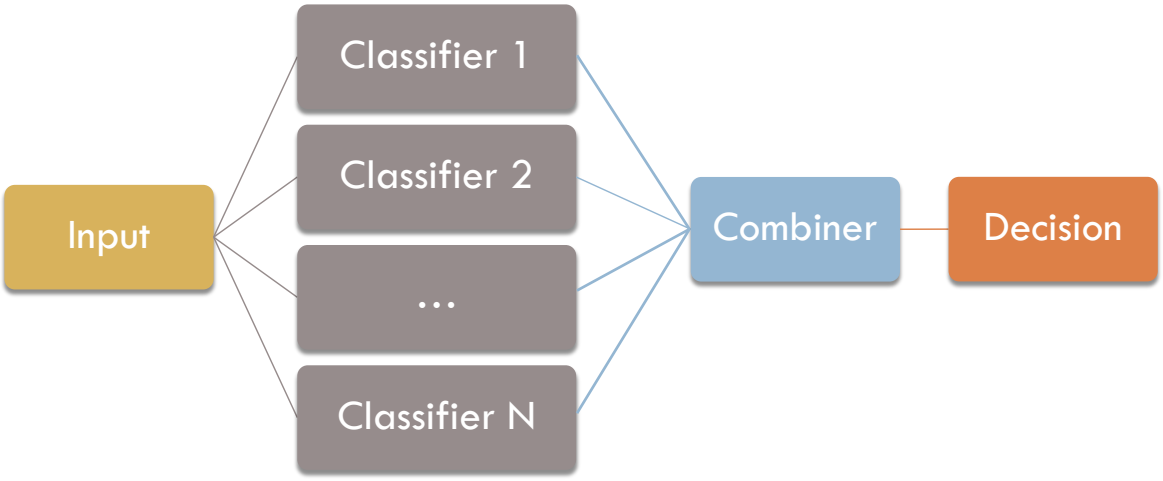

Figure 3.1: An example of parallel classifier ensemble architecture where N independent classifiers provide predictions which are then fused using an ensemble combination method.

3.1.1 Classifier Ensembles

Classifier ensembles are a way of combining different classifiers or experts with the goal of improving output accuracy through enhanced decision making. They have been applied to a wide range of real-world problems and shown to achieve better results compared to single-classifier methods (Oza and Tumer, 2008). Through aggregating the outputs of multiple classifiers in some way, their outputs are generally considered to be more robust. Ensemble methods continue to receive increasing attention from researchers and remain a focus of machine learning research (Woźniak et al., 2014; Kuncheva and Rodríguez, 2014).

Such ensemble-based systems often use a parallel architecture, as illustrated in Figure 3.1, where the classifiers are run independently and their outputs are aggregated using a fusion method. Other, more sophisticated, ensemble methods that rely on meta-learning may employ a stacked architecture where the output from a first set of classifiers is fed into a second level meta-classifier and so on.

The first part of creating an ensemble is generating the individual classifiers. Various methods for creating these ensemble elements have been proposed. These involve using different algorithms, parameters or feature types; applying different preprocessing or feature scaling methods; and varying (*e.g.* distorting or resampling) the training data.

For example, *Bagging* (bootstrap aggregating) is a commonly used method for ensemble generation (Breiman, 1996) that can create multiple base classifiers. It works by creating multiple bootstrap training sets from the original training data and a separate classifier is trained from each one of these sets. The generated classifiers are said to be diverse because each training set is created by sampling with replacement and contains a random subset of the original data. *Boosting* (*e.g.* with the AdaBoost algorithm) is another method where the base models are created with different weight distributions over the training data with the aim of assigning higher weights to training instances that are misclassified (Freund and Schapire, 1996).

The second part of ensemble design is choosing a combination rule to aggregate the outputs from the various learners, this is discussed in the next section.

3.1.2 Ensemble Combination Methods

Once it has been decided how the set of base classifiers will be generated, selecting the classifier combination method is the next fundamental design question in ensemble construction.

The answer to this question depends on what output is available from the individual classifiers. Some combination methods are designed to work with class labels, assuming that each learner outputs a single class label prediction for each data point. Other methods are designed to work with class-based continuous output, requiring that for each instance every classifier provides a measure of confidence[2] for each class label. These outputs may correspond to probabilities for each class and consequently sum to 1 over all the classes. That will be the case for the classifiers we will work with.

Although a number of different fusion methods have been proposed and tested, there is no single dominant method (Polikar, 2006). The performance of these methods is influenced by the nature of the problem and available training data, the size of the ensemble, the base classifiers used and the diversity between their outputs.

The selection of this method is often done empirically. Many researchers have compared and contrasted the performance of combiners on different problems, and most of these studies – both empirical and theoretical – do not reach a definitive conclusion (Kuncheva, 2014, p 178).

In the same spirit, we experiment with several information fusion methods which have been widely discussed in the machine learning literature. Our selected methods are described below; the descriptions generally assume that the individual classifiers have probabilistic outputs. A variety of other methods exist and the interested reader can refer to the thorough exposition by Polikar (2006).

3.1.2.1 Plurality voting

Each classifier votes for a single class label. The votes are tallied and the label with the highest number[3] of votes wins. Ties are broken arbitrarily. This voting method is very simple and does not have any parameters to tune. An extensive analysis of this method and its theoretical underpinnings can be found in Kuncheva (2004, p. 112).

3.1.2.2 Mean Probability Rule

The probability estimates for each class are added together and the class label with the highest average probability is the winner. This is illustrated in Figure 3.2. This is equivalent to the probability sum combiner which does not require calculating the average for each class. An important aspect of using probability outputs in this way is that a classifier's support for the true class label is taken into account, even when it is not the predicted label (*e.g.* it could have the second highest probability). This method has been shown to work well on a wide range of problems and, in general, it is considered to be simple, intuitive, stable (Kuncheva, 2014, p. 155) and resilient to estimation errors (Kittler et al., 1998) making it one of the more robust combiners discussed in the literature.

3.1.2.3 Median Probability Rule

Given that the mean probability used in the above rule is sensitive to outliers, an alternative is to use the median as a more robust estimate of the mean (Kittler et al., 1998). Under this rule each

[2] *e.g.* an estimate of the posterior probability for the class label.

[3] This differs with a *majority* voting combiner where a label must obtain over 50% of the votes to win. However, the names are sometimes used interchangeably.

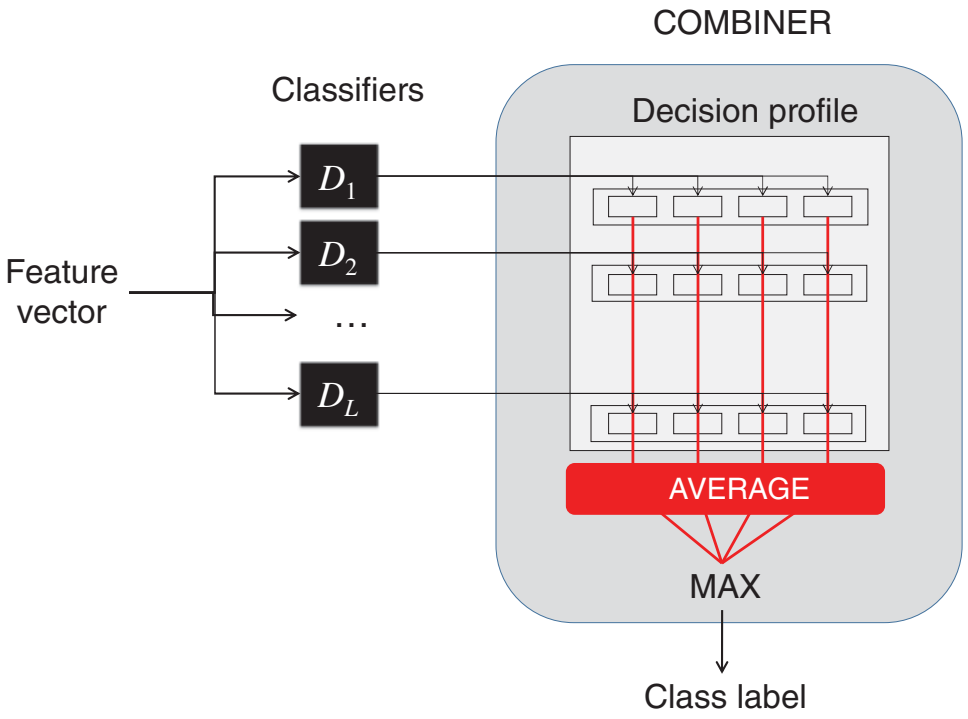

Figure 3.2: An example of a mean probability combiner. The feature vector for a sample is input to L different classifiers, each of which output a vector of confidence probabilities for each possible class label. These vectors are combined to form the decision profile for the instance which is used to calculate the average support given to each label. The label with the maximum support is then chosen as the prediction. Image reproduced from (Kuncheva, 2014, Fig. 5.5).

class label's estimates are sorted and the median value is selected as the final score for that label. The label with the highest median value is picked as the winner. As with the mean combiner, this method measures the central tendency of support for each label as a means of reaching a consensus decision.

3.1.2.4 Product Rule

For each class label, all of the probability estimates are multiplied together to create the label's final estimate (Polikar, 2006, p. 37). The label with the highest estimate is selected. This rule can theoretically provide the best overall estimate of posterior probability for a label, assuming that the individual estimates are accurate. A trade-off here is that this method is very sensitive to low probabilities: a single low score for a label from any classifier will essentially eliminate that class label.

3.1.2.5 Highest Confidence

In this simple method, the class label that receives the vote with the largest degree of confidence is selected as the final prediction (Kuncheva, 2014, p. 150). In contrast to the previous methods, this combiner disregards the consensus opinion and instead picks the prediction of the expert with the highest degree of confidence.

3.1.2.6 Borda Count

This method works by using each classifier's confidence estimates to create a ranked list of the class labels in order of preference, with the predicted label at rank 1. The winning label is then selected using the Borda count[4] algorithm (Ho et al., 1994). The algorithm works by assigning points to labels based on their ranks. If there are N different labels, then each classifier's preferences are assigned points as follows: the top-ranked label receives N points, the second place label receives $N-1$ points, third place receives $N-2$ points and so on with the last preference receiving a single point. These points are then tallied to select the winner with the highest score.

The most obvious advantage of this method is that it takes into account all of each classifier's preferences, making it possible for a label to win even if another label received the majority of the first preference votes.

3.1.3 Preliminary Experiments

We approached the design of our system by conducting some preliminary experiments on the TOEFL11-TRAIN and TOEFL11-DEV data, described in §2.3.1.4, to guide our system development. In this section we also outline some of these experiments and their results.

Our overall approach in terms of features and classifiers used is a fairly standard one. One difference from most approaches, but inspired by Tetreault et al. (2012), is that we train multiple classifiers over subsets of the features, over different feature representations, and over different regularisation approaches; we then combine them in ensembles.

3.1.3.1 Classifiers

For our machine learning methods, we experimented with both Support Vector Machine (SVM) classifiers and Logistic Regression classifiers. Post hoc analysis of the shared task systems, detailed in §2.3.3.3, shows that these two classifier types were the most popular among the entries. We specifically use the LIBLINEAR software package (Fan et al., 2008),[5] which is well-suited to text classification tasks with large numbers of features and large numbers of documents. LIBLINEAR provides both logistic regression and linear SVMs. Although both methods perform work with high-dimensional input, as we described in §2.1.3, there are some differences that we must consider in choosing which one to use in different scenarios.

Linear SVM Classifier SVMs are inherently binary classifiers and a common way to adapt them for multi-class problems is through a one-vs-all (OVA) approach.[6] Another common alternative is a one-vs-one (OVO) method that builds $\frac{N(N-1)}{2}$ binary classifiers for all pairwise combinations. It has been found that the OVR approach works best in NLI (Brooke and Hirst, 2012b), and our experiments confirmed this and we therefore adopt this approach. Additionally, an SVM is a margin-based classifier and does not output probability estimates for each class label, although there are additional methods to map the outputs to probabilities (Platt, 2000).

[4] This method is generally attributed to Jean-Charles de Borda (1733–1799), but evidence suggests that it was also proposed by Ramon Llull (1232–1315).

[5] Available from http://www.csie.ntu.edu.tw/%7Ecjlin/liblinear/

[6] Also known as a one-vs-rest (OVR) approach.

Logistic Regression Classifiers This algorithm is inherently multi-class, meaning that OVA and OVO approaches are not required. The logistic regression classifier is also probabilistic and provides continuous probability estimates for each class label, which are required by most of our combination methods (§3.1.2). We therefore follow the methodology of Tetreault et al. (2012) and use the logistic regression classifiers in the work in this thesis chapter.

In both single feature experiments and the ensemble we train each model on one of our feature types, which we describe in the next section.

3.1.3.2 Feature Set

We mostly use standard feature types in our system. Our feature set includes:

Character n-grams As described in §2.3.2.2. We limit our n-grams to the order of $n = 1$–3 and the were extracted from within word boundaries.

Function Word n-grams In addition to the standard function word unigram features we described in §2.3.2.5 we also propose an extension to higher order n-grams, which will be defined next in §3.1.3.4. For extracting the features, a standard English function word list was obtained from the Onix Text Retrieval Toolkit.[7]

Word n-grams We used this standard feature (§2.3.2.3), with $n = 1$–3. The n-grams included punctuation marks.

POS n-grams This purely syntactic feature (defined in §2.3.2.6) was also used. In our work, the POS tags for each text are predicted with a POS tagger and n-grams of order 1–3 are extracted from the tags. These n-grams capture (very local) syntactic patterns of language use and are used as classification features. Previous research and results from our own experiments show that sequences of size 4 or greater achieve lower accuracy, possibly due to data sparsity, so we do not generally present them in our work. Additionally, we also compare the performance of this feature using different POS tagsets, as we describe below in §3.1.3.5.

TSG fragments We described this relatively new feature in §2.3.2.8. Here, as an approximation to deploying the Bayesian approach to induce a TSG, each text is first parsed to obtain constituent parse trees. TSG fragments are then extracted from the parse trees using the TSG system made available by Post and Gildea (2009).[8]

Stanford Dependencies We also evaluate the grammatical dependency features we described in §2.3.2.9. Here we extract all the basic (rather than the collapsed) dependencies returned by the Stanford Parser (de Marneffe et al., 2006) and also apply the POS transformation.

Adaptor Grammar n-grams As outlined in §2.3.2.10, this feature captures n-grams of arbitrary length. We use the mixed POS and function word variant, with the features being extracted using the exact method of Wong et al. (2012).

[7] `http://www.lextek.com/manuals/onix/stopwords1.html`
[8] `https://github.com/mjpost/dptsg`

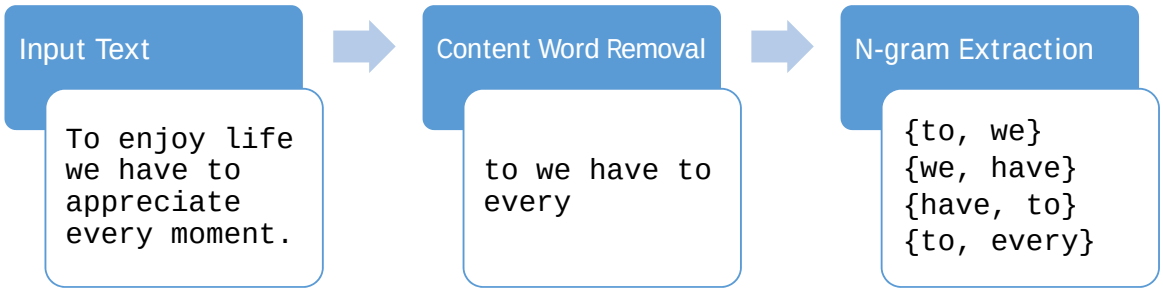

Figure 3.3: An example of how function word bigrams are extracted: content words are first stripped from the input and n-grams are extracted from the remaining tokens.

N	Accuracy
1	51.38
2	**59.73**
3	52.14

Table 3.1: NLI results for Function Word n-grams of order N. Our proposed Function Word bigram and trigram features outperform the commonly used unigrams.

3.1.3.3 Feature Representation

Most NLI studies have used two types of feature representations: binary (presence or absence of a feature in a text) and normalized frequencies. Although binary feature values have been used in some studies (*e.g.* Wong and Dras (2011)), most have used frequency-based values. In our system we experiment with both types.

3.1.3.4 Function Word n-grams

We did devise and test a new feature that attempts to capture patterns of function word use at the sentence level. This is because function words have been found to be an important feature in stylistic classification and an extension to n-grams could further improve accuracy.

We define function word n-grams as a type of word n-grams where content words are skipped: they are thus a specific subtype of skip-gram discussed by Guthrie et al. (2006). For example, the sentence *We should all start taking the bus* would first be reduced to *we should all the* by removing the content words, from which we would then extract the n-grams. This process is also demonstrated in Figure 3.3 with a different sentence. We test bigram and trigram versions of this feature.

To see if this sort of feature is at all useful, we conduct a preliminary experiment to compare the performance of the widely used unigrams against the proposed bigram and trigram variants, which have not been previously used for NLI. We trained a logistic regression classifier for each feature type using the TOEFL11-TRAIN and TOEFL11-DEV data, evaluating them under 10-fold cross-validation. The results are listed in Table 3.1. They show that function word skip-grams are more informative than the simple function word counts that have been previously used. We also observe that performance drop with trigrams and therefore we do not include them in further experiments.

```
I_PRP would_MD n't_RB agree_VB with_IN this_DT statement_NN ._.

I_PPIS1 would_VM n't_XX agree_VV0 with_IW this_DD1 statement_NN1 ._.

***

This_DT is_VBZ can_MD just_RB be_VB achieved_VBN by_IN concentrating_VBG
and_CC working_VBG hard_RB on_IN one_CD subject_NN ._.

This_DD1 is_VBZ can_VM just_RR be_VB0 achieved_VVN by_II concentrating_VVG
and_CC working_VVG hard_RR on_II one_MC1 subject_NN1 ._.

***

A_DT reason_NN for_IN this_DT could_MD be_VB also_RB a_DT problem_NN
of_IN teaching_NN skills_NNS ._.

A_AT1 reason_NN1 for_IF this_DD1 could_VM be_VB0 also_RR a_AT1 problem_NN1
of_IO teaching_NN1 skills_NN2 ._.
```

Figure 3.4: Three example sentences tagged with both the PTB (red) and CLAWS2 (blue) tagsets.

3.1.3.5 Comparing POS Tagsets

In extracting POS n-gram features from the data we were faced with a choice between two automatic taggers: The Stanford POS tagger and the Robust Accurate Statistical Parsing (RASP) system.[9]

While both systems are trained on large datasets and considered to be accurate, a key difference is the tagset that they use. The Stanford system uses the Penn Treebank (PTB) tagset which has 36 tags. RASP uses a modified version of the CLAWS2 tagset[10] which has 166 tags. To demonstrate some of these differences, three example sentences tagged with both tagsets are shown in Figure 3.4. Some of the notable differences include the more linguistically fine-grained categorization of pronouns,[11] determiners and prepositions in the CLAWS2 tagset. For example, the prepositions *with*, *by*, *for* and *of* are all given the same PTB tag while CLAWS2 assigns them each a unique tag. Additionally, the CLAWS2 tagset also distinguishes open-class words more finely, for example, through verb sub-categorization.

Given the large difference in the number of tags they use, we decided to directly compare both variants to gauge their performance for this task.

To compare the tagsets we trained individual logistic regression classifiers for n-grams of order 1–4 using both tagsets and tested them. This was done by training the TOEFL11-TRAIN and TOEFL11-DEV data and evaluating them under 10-fold cross-validation. The results are listed in Table 3.2 and show that the CLAWS2-tagged data provided better performance in all cases. While it is possible that CLAWS2's better performance could be attributed to other factors such as tagging accuracy, we do not believe this to be the case as the Stanford Tagger is known for its high accuracy.[12]

These differences are quite clear and accordingly, we use the features extracted via the RASP tagger in our experiments. This finding also has implications for other syntactic features that make use of POS tags, such as Adaptor Grammars, Stanford Dependencies and Tree Substitution Grammars.

[9] `http://ilexir.co.uk/applications/rasp/`
[10] `http://ucrel.lancs.ac.uk/claws2tags.html`
[11] CLAWS2 has over 20 pronoun tags.
[12] 97% on the standard WSJ22-24 test set, see `http://nlp.stanford.edu/software/pos-tagger-faq.shtml`

N	PTB	CLAWS2
1	34.03	43.76
2	48.85	58.93
3	51.06	**59.39**
4	49.85	52.81

Table 3.2: Classification accuracy results for POS n-grams of order N using both the PTB and CLAWS2 tagsets. The larger CLAWS2 tagset used by RASP tagger performed substantially better for all values of N.

Feature	Train + Dev Set	Test Set
Chance Baseline	9.1	9.1
Character unigram	33.99	34.70
Character bigram	51.64	49.80
Character trigram	66.43	66.70
RASP POS unigram	43.76	45.10
RASP POS bigram	58.93	61.60
RASP POS trigram	59.39	62.70
Function word unigram	51.38	54.00
Function word bigram	59.73	63.00
Word unigram	74.61	75.50
Word bigram	74.46	76.00
Word trigram	63.60	65.00
TSG Fragments	72.16	72.70
Stanford Dependencies	73.78	75.90
Adaptor Grammar POS/FW n-grams	69.76	70.00

Table 3.3: Classification results for our individual features.

This issue of tagset size is further explored in greater depth in Chapter 5, as part of the experiment described in §5.8.

3.1.3.6 Individual Feature Performance

We next ran some similar tests to evaluate the performance of our other feature types. This was done by training a single classifier on the TOEFL11-TRAIN and TOEFL11-DEV data for each feature type and evaluating them under 10-fold cross-validation. Additionally, we also trained on the combined TOEFL11-TRAIN and TOEFL11-DEV data and evaluated the features on the held-out TOEFL11-TEST set once the data labels were released after the shared task. The results are summarized in Table 3.3.

Character n-grams are an informative feature and our results are very similar to those reported by previous researchers (Tsur and Rappoport, 2007). In particular, it should be noted that the use of punctuation is a very powerful feature for distinguishing languages. Romance language speakers were most likely to use more punctuation symbols (colons, semicolons, ellipsis, parenthesis, etc.) and at higher rates. Chinese, Japanese and Korean speakers were far less likely to use punctuation in their writing.

The performance of word n-grams, TSG fragments, adaptor grammar n-grams and Stanford Dependencies was very strong and comparable to previously reported research. Word n-grams do particularly well, in spite of the topic balanced data.

Hapax Legomena and Dis Legomena Given the strong performance of the lexical features, we also investigated word unigrams at a more fine-grained level. The special word categories *Hapax Legomena* and *Dis legomena* refer to words that appear only once and twice, respectively, in a complete text. In practice, these features are a subset of our word unigram feature, where *Hapax Legomena* correspond to unigrams with a frequency of 1 within a text and *Dis legomena* are unigrams with a frequency of 2. In our experimental results we found that *Hapax Legomena* alone provides an accuracy of 61%. Combining the two features together yields an accuracy of 67%. This is an interesting finding as both of these features alone provide an accuracy close to the whole set of word unigrams. They were, however, not included in our system due to their lower performance.

3.1.3.7 Ensemble Construction and Evaluation

To construct our ensemble, we train individual logistic regression classifiers on a single feature type (*e.g.* POS n-grams), using a specific feature value representation and classifier. We utilize a parallel ensemble structure where the classifiers are run on the input texts independently and their results are then fused into the final output using a combiner.

In the course of our experiments we have observed that the effect of the feature representation varies with the feature type, size of the feature space and the learning algorithm itself. Based on this, then, we decided to generate two distinct classifiers for each feature type, one trained with frequency-based values (raw counts scaled using the l^2-norm) and the other with binary. This results in two base classifiers per feature type (*e.g.* POS2grams-bin and POS2grams-freq). Our experiments also assess both their individual and joint performance.

Furthermore, in our preliminary experiments we also observed that some feature types performed better with L1-regularization and others with L2-regularization. As a result of this we also extend our ensemble by generating classifiers using both regularization methods and evaluate their individual and combined performance. This generates an ensemble with four classifiers per feature type, (*e.g.* POS2grams-bin-L1, POS2grams-freq-L1, POS2grams-bin-L2 and POS2grams-freq-L2).

Bagging Additionally, we also experiment with bagging (bootstrap aggregating), a commonly used method for ensemble generation (Breiman, 1996) to generate multiple classifiers per feature type. Our experiments here did not find any improvements in accuracy, even with larger numbers of bootstrap samples (50 or more). Bagging is said to be more suitable for unstable classifiers which have greater variability in their performance[13] and are more susceptible to noise in the training data (Breiman, 1996). In our experiments with individual feature types we have found the classifiers to be quite stable in their performance, across different folds and training set sizes. This is one potential reason why bagging did not yield significant improvements. Accordingly, we did not investigate this any further and did not use the method for our submission.

Combination Methods Of the methods outlined in §3.1.2 we found the mean probability combiner to be the best performing and our results are reported using this combiner. However, we also note that

[13] *i.e.* when changing the training data results in significant differences in model performance.

Ensemble Components	bin-L1	bin-L2	freq-L1	freq-L2	CV	Test set
(1) Complete Ensemble	x	x	x	x	81.50	81.60
(2) Only binary values	x	x			**82.46**	**83.10**
(3) Only freq values			x	x	65.28	67.20
(4) L1-regularized only	x		x		80.33	81.10
(5) L2-regularized only		x		x	81.42	81.10
(6) Bin, L1-regularized only	x				81.57	82.00
(7) Bin, L2-regularized only		x			82.00	82.50

Table 3.4: NLI classification results for our ensembles, best result in column in bold (binary values with L1- and L2-regularized solvers).

the differences across all combination methods was roughly 1–2%. Any new approach to ensemble combination here would consequently want to be radically different to expect a notable improvement in performance.

Ensemble Results We used the same training and testing data as the previous section (§3.1.3.6). Table 3.4 shows the results from our ensembles. The feature types included in the ensemble are those whose results are listed individually in Table 3.3. (So, for example, we only use the RASP-tagged POS n-grams, not the Penn Treebank ones.) The complete ensemble consists of four classifiers per feature type: L1- and L2-regularized versions with both binary and frequency values.

Combining Regularization Approaches Our results show that combining both the L1- and L2-regularized classifiers in the ensemble provided a small increase in accuracy. Ensembles with only one of the L1 or L2-regularization terms have lower accuracy than the combined methods (*e.g.* comparing row 2 of Table 3.4 with rows 6 and 7).

3.1.3.8 Proficiency-level Based Classification

Our final preliminary experiment examined how we could utilize the proficiency level information provided in the TOEFL11 corpus (texts are marked as either low, medium or high proficiency) in our ensembles. We did this by evaluating classifiers that are trained using only texts from specific proficiencies.

Tetreault et al. (2012) established that the classification accuracy of their system varied across proficiency levels, with high proficiency texts being the hardest to classify. This is most likely due to the fact that writers at differing skill levels commit distinct types of errors at different rates (Ortega, 2009, for example). If learners of different backgrounds commit these errors with different distributions, these patterns could be used by a learner to further improve classification accuracy.

We evaluated this by segregating the TOEFL11-TRAIN and TOEFL11-DEV sets into subsets of different proficiency levels. We then trained logistic regression models on different subsets of TOEFL11-TRAIN and tested on different subsets of TOEFL11-DEV, aiming to find patterns about which training subsets were best suited for classifying which test subsets.

Table 3.5 shows our results for this experiment, listing the different training and testing subsets. The numbers show that in general texts should be classified with a learner trained with texts of a sim-

Training Set	Dev Set	Accuracy
Low	Low	52.2
Medium	Low	72.1
High	Low	40.3
Low + Medium	Low	**76.0**
All	Low	75.2
Low	Medium	40.7
Medium	Medium	83.6
High	Medium	62.1
Medium + High	Medium	85.3
Low + Medium	Medium	83.8
All	Medium	**86.8**
Low	High	16.1
Medium	High	68.1
High	High	65.7
Medium + High	High	74.7
All	High	**75.2**

Table 3.5: Results for classifying the test set documents using classifiers trained with a specific proficiency level. Each level's best result in bold.

ilar proficiency. They also show that not all texts in a proficiency level are of uniform quality as some levels perform better with data from the closest neighbouring levels (*e.g.* Medium texts perform best with data from all proficiencies), suggesting that the three levels form a larger proficiency continuum where users may fall in the higher or lower ends of a level. A larger corpus with more fine-grained proficiency categories (*i.e.* more than three levels) could help address this and potentially achieve higher results. We also note that training on low proficiency data for classifying high proficiency texts, and vice versa, produces very low performance.

Based on these results, we decided to use this approach in one of our shared task entries. This was made possible as the unlabelled version of the TOEFL11-TEST set released during the shared task included the proficiency information. A proficiency classifier would be required in other practical scenarios where the proficiency details of test documents are unavailable.

3.1.4 Shared Task Systems and Results

Based on the preliminary experiments described above in §3.1.3, we created the following five systems for entry into the shared task.

1. A system based on the full ensemble of four classifiers per feature type. This, and the other four ensembles below are all combined with the mean probability fusion rule.
2. An ensemble of L1-regularized classifiers using only binary feature representations.
3. An ensemble of L2-regularized classifiers using only binary feature representations.
4. An ensemble of L1- and L2-regularized classifiers using only binary feature representations.
5. A proficiency-segregated version of the full ensemble, with the low proficiency texts classified by models trained on low and medium proficiency texts and medium and high proficiency texts being classified by models trained on all data.

System	Accuracy (%)
1	79.6
2	79.7
3	79.7
4	**80.1**
5	79.5

Table 3.6: Official shared task results for our 5 systems.

		Predicted										
		ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
Actual	ARA	**[78]**	0	5	2	4	1	1	0	5	4	0
	CHI	1	**[89]**	1	1	1	0	4	2	1	0	0
	FRE	1	1	**[86]**	4	3	3	0	0	2	0	0
	GER	0	1	2	**[92]**	1	0	1	0	1	0	2
	HIN	1	2	0	0	**[76]**	0	0	0	2	18	1
	ITA	0	0	5	3	0	**[88]**	0	0	2	0	2
	JPN	1	6	1	2	1	0	**[83]**	3	0	0	3
	KOR	1	9	4	1	1	0	10	**[71]**	2	1	0
	SPA	5	1	10	1	2	10	0	3	**[64]**	0	4
	TEL	2	1	0	1	17	0	0	0	0	**[79]**	0
	TUR	5	5	4	1	0	2	2	3	3	0	**[75]**

Table 3.7: Confusion matrix of our best system in the 2013 NLI Shared Task.

The official results for all of our systems are shown in Table 3.6. All systems had very similar results, with System 4 achieving the best results. This shows that combining different regularization approaches can help improve accuracy.

We also note that the proficiency-segregated ensemble that used separate models for low proficiency texts failed to achieve any improvement on the test set. The results are almost identical to the non-segregated version (System 1). This is not entirely surprising given that the medium and high proficiency categories make up some 89% of the TOEFL11 data, as we described in §2.3.1.4.

The results listed here are those reported by the shared task organizers. The systems we submitted were trained on only the TOEFL11-TRAIN set and we had not included the TOEFL11-DEV set like other teams. When trained on both of these sets the best result on the test set improves to 83.1%, 3% higher than our best submission result, and close to the winning system's result (83.6%).

The confusion matrix of the output from our best system is presented in Figure 3.5. A tabulated version with the exact counts is also presented in Table 3.7. We see that German is the most correctly classified class while Spanish has the worst performance due to confusion with two other Romance languages: French and Italian. Interestingly, however, French and Italian are not often confused with Spanish. A similar pattern is also observed for Korean being confused for Japanese and Chinese. Finally, we also note that the greatest degree of confusion is between Hindi and Telugu, as we discuss in the next section.

A plot of each team's best system was previously shown in Figure 2.12, with our results listed as "MQ". There was no statistically significant difference between the top five teams.

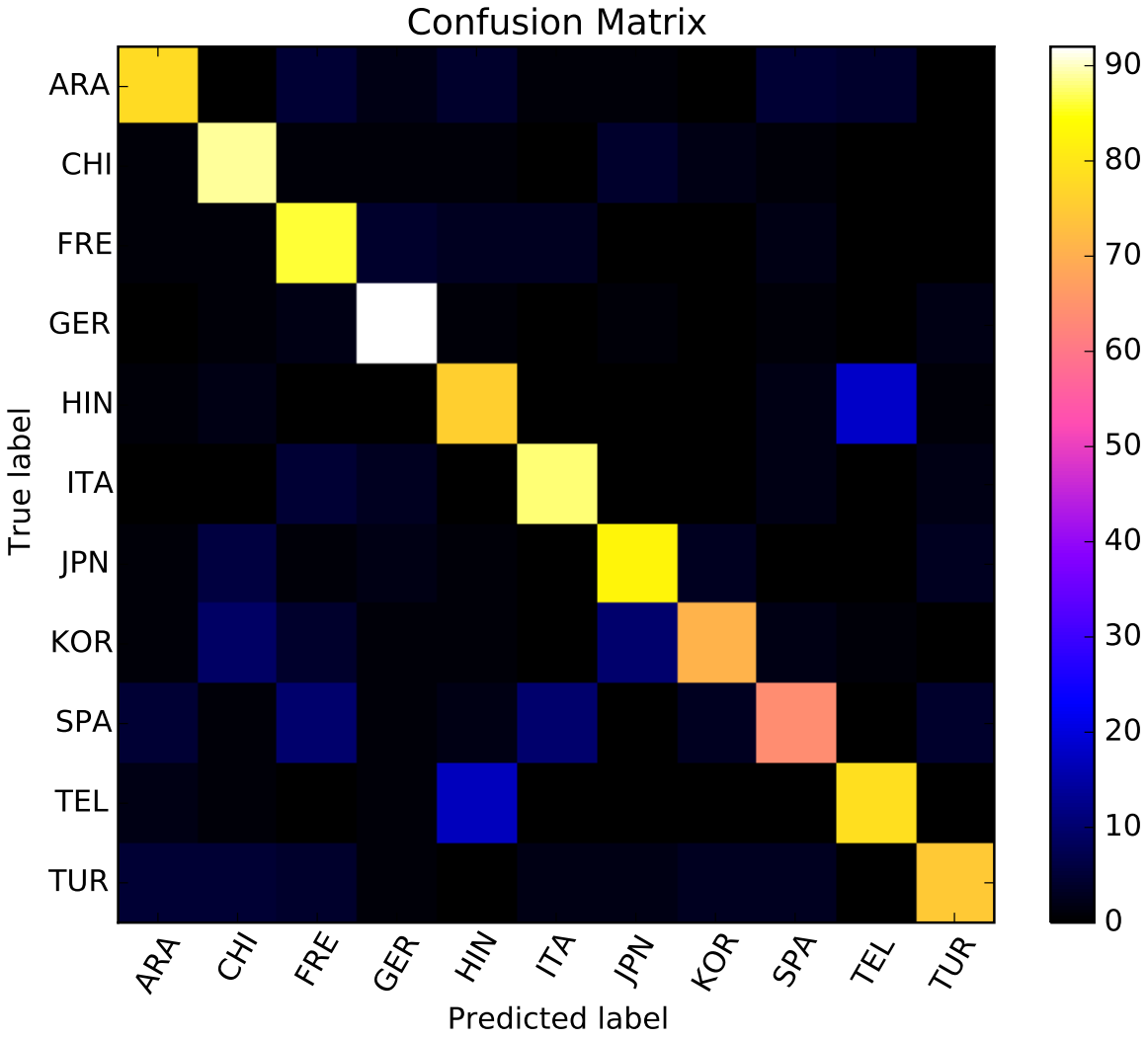

Figure 3.5: Confusion matrix of our best system in the 2013 NLI Shared Task.

3.1.5 Discussion

Our experiments for building our shared task entry resulted in a number of useful results. For the most part we focused on evaluating the different tagsets and building ensembles. We demonstrated that a more linguistically fine-grained tagset achieves higher classification results than a more coarse one, an issue which we will revisit more thoroughly in §5.8. We also demonstrated that classifier ensembles are useful for this task; the results achieved by our ensemble were higher than any single feature could achieve alone. Additionally, our results indicate that combining different machine learners in an ensemble can be useful.

We also tested a proficiency-segregated ensemble, although this did not result in improved performance. Several other shared task entries also attempted to use the proficiency level meta-data included with the corpus to aid classification, *e.g.* the systems by Kyle et al. (2013) and Brooke and Hirst (2013), although none of them were able to successfully leverage this information to improve their system's accuracy.

In terms of feature engineering, we proposed the use of function word bigrams as a new feature and also reported results suggesting that using larger POS tagsets can improve NLI accuracy.

3.1.5.1 Feature Analysis

We conducted a brief analysis of our extracted features, looking at the most predictive ones according to their Information Gain. Although we did not find any obvious indicators of topic bias, we noted some other issues of potential concern.

Chinese, Japanese and Korean speakers make excessive use of phrases such as *however*, *first of all* and *secondly*. At first glance, the usage rate of these phrases seems unnaturally high (more than 50% of Korean texts had a sentence beginning with *however*). This could perhaps be a cohort effect relating to those individually attempting this particular TOEFL exam, rather than an L1 effect: it would be useful to know how much variability there is in terms of where candidates come from.

It was also noticed that many writers mention the name of their country in their texts, and this could potentially create a high correlation between those words and the language class label, leading perhaps to an artificial boosting of results. For example, the words *India*, *Turkey*, *Japan*, *Korea* and *Germany* appear with high frequency in the texts of their corresponding L1 speakers — hundreds of times, in fact, in contrast to frequencies in the single figures for speakers of other L1s. These might also be an artefact of the type of text, rather than related to the L1 as such.

Although the new topic-balanced TOEFL11 data is a marked improvement over previously used corpora, these features highlight an important issue: other categories of non-topical words, such as the toponyms we observed here, can exert undue influence on results based on lexical features. The models used to obtain such results must be inspected in order to better understand the types of features that are highly weighted by the model. This is an issue that we will revisit in §4.2 and §6.3.

3.1.5.2 Hindi vs. Telugu

We single out here this language pair because of the high level of confusion between the two classes. Looking at the results obtained by other teams, we observe that this language pair provided the worst classification accuracy for almost all teams. No system was able to achieve an accuracy of 80% for Hindi (something many achieved for other languages). In analysing the actual and predicted

Each/DT person/NN should/MD **has/VBZ** one/CD specialized/VBN subject/NN ./.

As/IN we/PRP have/VBP noticed/VBN ,/, societies/NNS must/MD **consists/VBZ** of/IN people/NNS of/IN different/JJ **knowleges/NNS** and/CC it/PRP is/VBZ impossible/JJ to/TO find/VB a/DT person/NN who/WP is/VBZ creative/JJ in/IN many/JJ academic/JJ fields/NNS ./.

It/PRP is/VBZ very/RB **intresting/JJ becouse/NN** before/IN you/PRP travel/VBP to/TO the/DT country/NN ,/, the/DT group/NN **giude/NN** has/VBZ has/VBZ **seted/VBN** the/DT **shcudual/NN** ./.

I/PRP both/DT agree/VBP and/CC **disgree/VBP** with/IN those/DT two/CD **statemant/JJ** ./.

from/IN my/PRP$ **prepactive/JJ** I/PRP **thik/VBP** what/WP all/DT matter/NN is/VBZ your/PRP$ **comfor/NN** what/WP ever/RB you/PRP feel/VBP **goood/NN** about/IN you/PRP choose/VBP ./.

in/IN the/DT **pass/NN cople/NN** years/NNS there/EX were/VBD **alot/NN** of/IN people/NNS that/WDT are/VBP successful/JJ ,/, and/CC so/RB much/JJ ambition/NN ./.

For/IN an/DT **exampl/NN** Bill/NNP Gaets/NNP **ang/NN** Inishtine/NNP ./.

they/PRP were/VBD so/RB **pore/RB** in/IN school/NN ,/, but/CC they/PRP were/VBD so/RB successful/JJ in/IN life/NN **infact/NN** they/PRP made/VBD the/DT World/NNP **chainged/VBD dy/JJ** two/CD **emprtant/JJ** things/NNS :/: well/RB and/CC ambition/NN ./.

Figure 3.6: POS tagging results for several learner sentences with grammatical and orthographic errors. Some of the errors are highlighted in bold. Results obtained from the Stanford POS tagger.

classes for all documents classified as Hindi and Telugu by our system, we find that generally all of the actual Hindi and Telugu texts (96% and 99%, respectively) are within the set. Our classifier is clearly having difficulty discriminating between these two specific classes.

Given this, we posit that the confounding influence may have more to do with the particular style of English that is spoken and taught within the country, rather than the specific L1 itself. Consulting other research about SLA differences in multi-lingual countries could shed further light on this.

Analysing highly informative features provides some clues about the influence of a common culture or national identity: in our classifier, the words *India*, *Indian* and *Hindu* were highly predictive of both Hindi and Telugu texts, but no other languages. Looking at individual language pairs in this way could lead to incremental improvement in the overall classification accuracy of NLI systems. We also look at this issue later in §4.1.4.

3.1.5.3 POS tagging learner texts

Many shared task systems, including ours, investigated using syntactic information such as POS tags as classification features. As we described, this is achieved by using automatic POS taggers based on statistical models, *e.g.* the Stanford POS Tagger, to automatically annotate the learner texts.

One issue to consider here is that the models used by these statistical taggers are trained on well-formed text from a standard variety of the language written by native speakers (*e.g.* news articles). When tested on such data, the models generally achieve high accuracies of 96% or higher. However, it cannot be assumed that these tools will achieve similar levels of accuracy on learner data, a distinct genre which they were not trained on. Furthermore, learner writing is known to contain various types of errors and ungrammatical sentences.

This is a consideration that has not gone unnoticed and several researchers have investigated this question. Van Rooy and Schäfer (2002) investigated this issue and report that "learner spelling errors contributed substantially to tagging errors", causing up to 38% of the POS tagging errors. Díaz-Negrillo et al. (2010) argue that the properties of learner language are systematically different from those assumed for the standard variety of the language and that this interlanguage cannot be considered a noisy variant of the native language. Instead of viewing this as a robustness issue, they suggest that a new POS model for learner language may be more suitable. Based on the results of their empirical analysis they highlight several issues with standard POS models and they propose a new tripartite POS annotated model that encodes properties based on the lexical stem, distribution and morphology. Another take on this issue can be found in Cahill (2015), which looks at the performance of syntactic parsers on text when errors are introduced. The recommendation here is that instead of adapting annotation schemes to accommodate learner errors, it may be better to shoehorn existing schemes to fit learner texts.

This evidence points to a performance degradation on learner data and suggests that the POS annotations used in many previous NLI studies are vulnerable to tagging errors. To investigate this we looked at the tagging results from the Stanford POS Tagger for several low proficiency texts from our TOEFL11 training data. A number of sentences with grammatical and orthographic errors from these texts, along with their tags, are shown in Figure 3.6. Some of the errors have been highlighted in bold. We observe several instances where spelling mistakes result in tagging errors (*e.g. statemant*/JJ) but we also note there are a number of other instances where misspelled words are tagged correctly, *e.g. emprtant*/JJ, *intresting*/JJ and *chainged*/VBD. This shows that the tagger does posses some degree of robustness against learner errors.

Although it is evident that statistical tagger performance is degraded on this data, it is unclear how this affects performance when the tags are used downstream for NLI. Evidence from previous work, and that presented throughout this thesis, suggests that they work well as features and have hence gained wide adoption. One possibility is that systematic learner errors result in systematic tagging errors. The NLI model can then learn to associate the tagging errors with the specific L1 groups that commit the writing errors that result in those particular erroneous tag patterns.

Of course it is also possible that the use of more accurate POS tags, obtained via more robust tagging or manual post-correction, could improve NLI results. This stance assumes that errors could reduce the efficacy of POS tag n-grams in distinguishing the different syntactic patterns used by different L1 groups. Testing this hypothesis will require carefully controlled experiments with sufficient data in order to reach a conclusive result. This line of inquiry, however, lies beyond the scope of this thesis and is left for future work.

3.1.5.4 SVM vs Logistic Regression

Post hoc analysis of the shared task results (§2.3.3.3) revealed that many of the top systems, including the winning entry, followed two broad patterns: they used SVM classifiers along with frequency-based feature values.

Our ensemble performance using logistic regression was consistent with that of previous research (Tetreault et al., 2012; Brooke and Hirst, 2012b) which concluded that there is a preference for binary feature values instead of frequency-based ones. Including both types in our ensemble did not improve results.

Given the better performance of SVM-based shared task systems we further investigated this by testing the performance of our individual features with a linear SVM classifier trained on normalized frequency feature values. The results for this comparison, listed in §3.2.2, showed substantially improved results over the logistic regression models. If output probability estimates for each class label are required for ensemble combination, these can be obtained by applying Platt scaling (Platt, 2000). Accordingly, we conducted the rest of the experiments in this thesis using such linear SVM models.

3.2 Measuring Feature Diversity

Results from our previous experiments show that while some feature types yield similar accuracies independently, such as those in Table 3.3, combining them can improve performance. This indicates that the information they capture is diverse, but how diverse are they and how can we measure the level of independence between the feature types?

This is a question that has not been tackled in NLI, despite researchers examining dozens of feature types to date. Research examining the combination of multiple features, *e.g.* Tetreault et al. (2012), has only look at the accuracy of individual features. In this experiment we examine one approach to measuring the degree of diversity between features.

An ablation study is a common approach in machine learning that aims to measure the contribution of each feature in a multi-component system. This ablative analysis is usually carried out by measuring the performance of the entire system with all components (*i.e.* features) and then progressively removing the components one at a time to see how the performance degrades.[14]

While useful for estimating the potential contribution of a component, this type of analysis does not directly inform us about the pairwise relation between any two given components. In their study of classifying discourse cohesion relations, Wellner et al. (2009) performed an ablation analysis of their feature classes and note:

> From the ablation results [...] it is clear that the utility of most of the individual features classes is lessened when all the other feature classes are taken into account. This indicates that multiple feature classes are responsible for providing evidence [about] given discourse relations. Removing a single feature class degrades performance, but only slightly, as the others can compensate. (Wellner et al. (2009, p. 122))

This highlights the need to quantify the overlap between any two given components in a system. Our approach to estimating the amount of diversity between two feature types is based on measuring the level of agreement between the two for predicting labels on the same set of documents. Here, we aim to examine feature differences by holding the classifier parameters and data constant.

Previous research has suggested that Yule's Q coefficient statistic (Yule, 1912; Warrens, 2008) is a useful measure of pairwise dependence between two classifiers (Kuncheva et al., 2003). This notion of dependence relates to complementarity and orthogonality, and is an important factor in combining classifiers (Lam, 2000).

Yule's Q statistic is a correlation coefficient for binary measurements and can be applied to classifier outputs for each data point where the output values represent correct (1) and incorrect (0)

[14] Other variations exist, *e.g.* Richardson et al. (2006) and Wellner et al. (2009).

predictions made by that learner. Each classifier C_i produces a result vector $y_i = [y_{i,1}, \cdots, y_{i,N}]$ for a set of N documents where $y_{i,j} = 1$ if C_i correctly classifies the j^{th} document, otherwise it is 0. Given these output vectors from two classifiers C_i and C_k, a 2×2 contingency table can be derived:

	C_k **Correct**	C_k **Wrong**
C_i **Correct**	N^{11}	N^{10}
C_i **Wrong**	N^{01}	N^{00}

Here N^{11} is the frequency of items that both classifiers predicted correctly, N^{00} where they were both wrong, and so on. The Q coefficient for the two classifiers can then be calculated as:

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}$$

This distribution-free association measure[15] is based on taking the products of the diagonal cell frequencies and calculating the ratio of their difference and sum.[16] Q ranges between −1 and +1, where −1 signifies negative association, 0 indicates no association (independence) and +1 means perfect positive correlation (dependence).

3.2.1 Methodology and Data

In §3.1 we concluded that linear SVM classifiers with frequency-based features produced the best results and discussed this in §3.1.5.4. Accordingly, that is the learner we use in this section (and the rest of the thesis). All of our classifiers here are always of the same type, a linear SVM classifier, but they are trained with different features on the very same dataset. This allows us to measure the dependency between feature types themselves.

We use the TOEFL11 data for this study. We use classification accuracy as our evaluation metric, obtained by training on the TOEFL11-TRAIN and TOEFL11-DEV sets and testing on the TOEFL11-TEST set. We take this approach, instead of cross-validation, because shared task results from our own work earlier in this chapter, and other shared task participants, has demonstrated that the performance of both approaches is very similar.

We employ the same set of syntactic and lexical features that we detailed in §3.1.3.2. These include: Adaptor Grammars (AG), character n-grams,[17] Function word unigrams and bigrams, Word and Lemma n-grams, CFG Production Rules, Penn Treebank (PTB) part-of-speech n-grams, RASP part-of-speech n-grams, Stanford Dependencies with POS transformations, and Tree Substitution Grammar (TSG) fragments.

3.2.2 Results

We begin by assessing the classification accuracy for each of our individual features, as shown in Figure 3.7. We observe that some feature pairs, such as Adaptor Grammars and Character trigrams and Lemma unigrams and Stanford Dependencies, achieve very similar accuracy.

[15] We also note that this is equivalent to the 2×2 version of Goodman and Kruskal's gamma measure for ordinal variables.

[16] Division by zero is possible here, see Bakeman and Quera (2011, p. 115) for more details.

[17] We restrict our investigation here to 1–3-grams, as in §3.1. Recent work has also shown improvements from longer sequences for certain feature types (Jarvis et al., 2013; Ionescu et al., 2014).

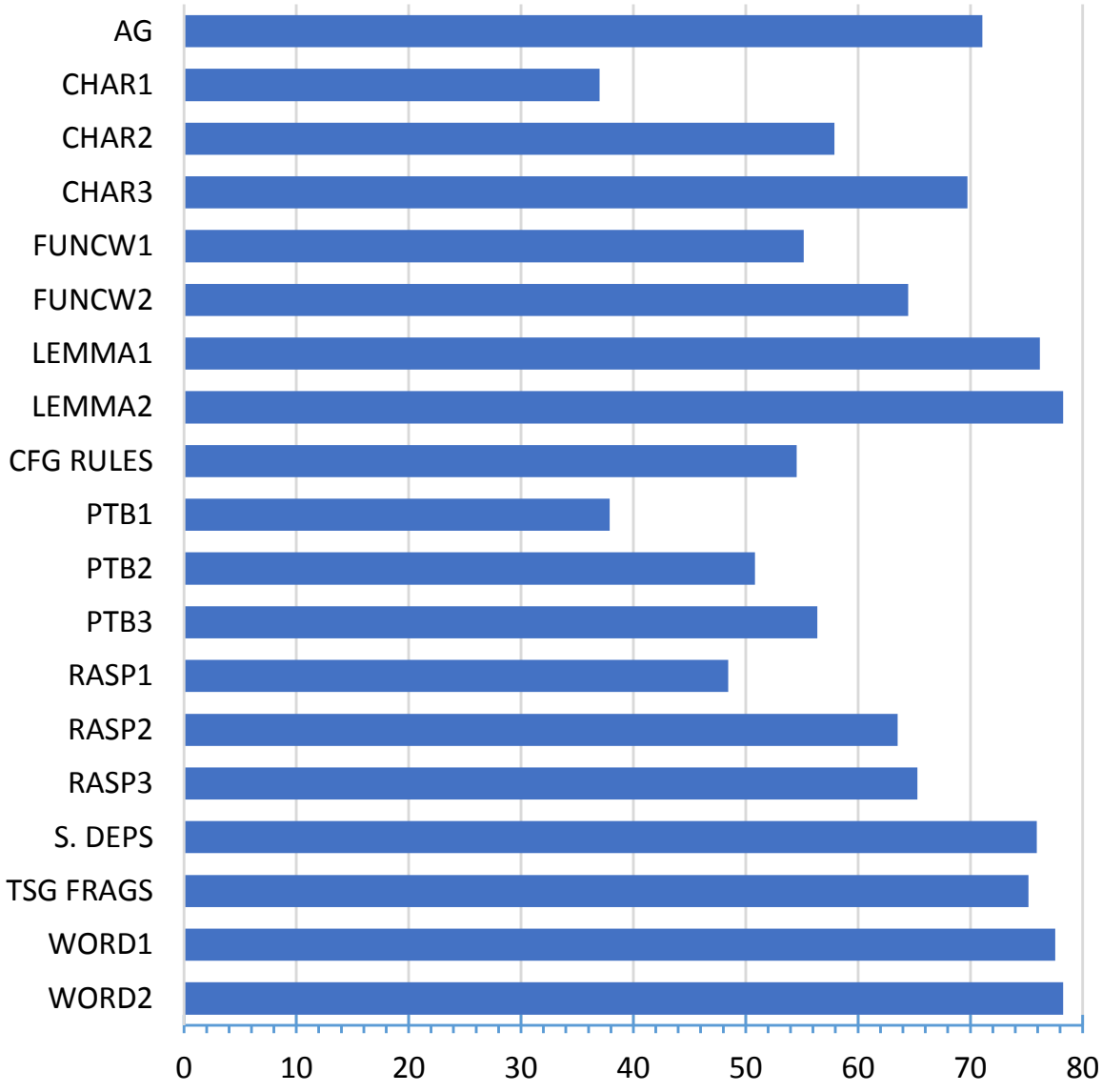

Figure 3.7: NLI accuracy per feature type on the TOEFL11 test set.

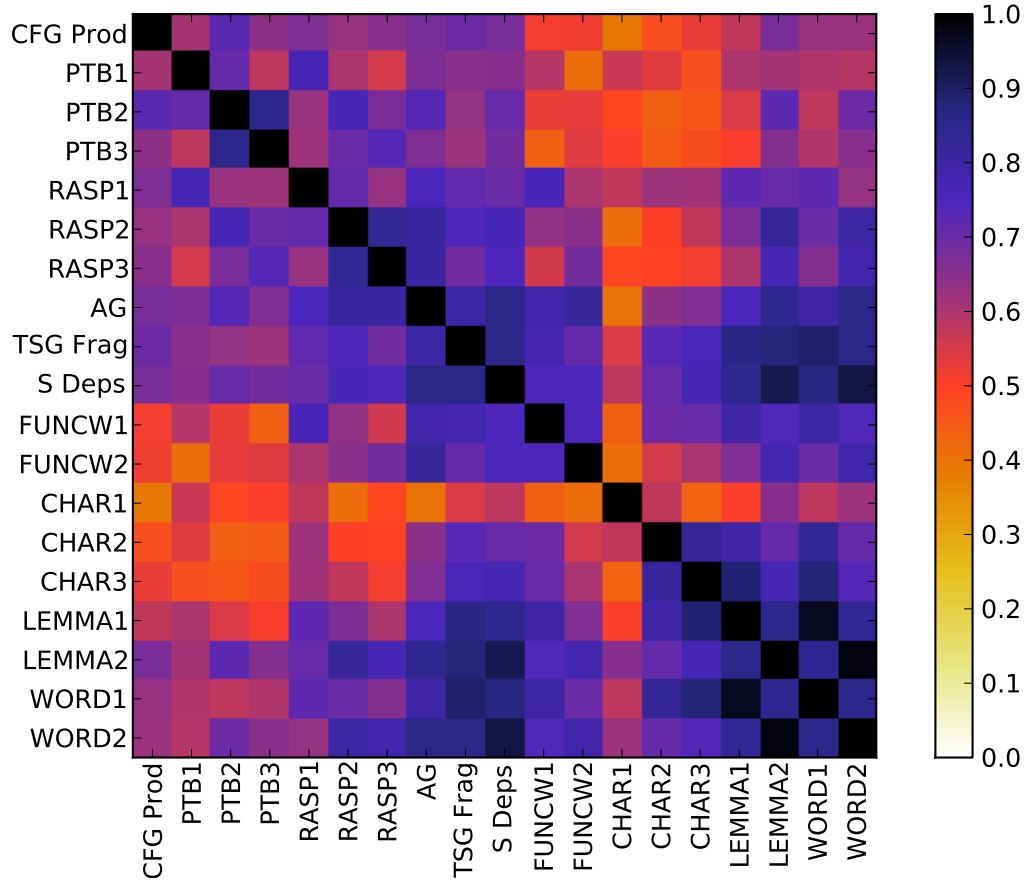

Figure 3.8: The Q-coefficient matrices of our feature set. The matrices are displayed as heat maps.

The matrix of the Q-coefficients for all features are shown graphically in Figure 3.8. This matrix is symmetric by definition of Q. The most discernible feature is the red cluster (representing the lowest Q-coefficient values) in the bottom left of the matrix. This region covers the correlations between syntactic and lexical features, showing that they differ the most.

Another interesting aspect is the strong correlations between the lexical features, shown by the clustering of high values in the bottom right corner. It also shows that character n-grams capture similar information to words. Even character unigrams – the lowest performing lexical feature – show much stronger dependence with word unigrams than other syntactic features. Additionally, the high values in the bottom middle section of the matrix show that Stanford Dependencies and TSG fragments largely capture the same information as Word and Lemma bigrams. These issues are explored further in §3.2.3.

In contrast to the lexical features, the syntactic ones show much lower inter-correlation levels, evidenced by lower values in the top left corner and absence of a visible cluster. This seems to indicate that there is greater diversity among these features.

Such analyses can help us better understand the linguistic properties of features and guide interpretation of the results. This knowledge can also be useful in creating classifier ensembles. One goal in creating such committee-based classifiers is the identification of the most diverse independent learners and this method can be applied to that end. To assess this, we also measure the classification accuracy for all 171 possible feature pair combinations f_i and f_j in our feature set.[18] Each pair is combined in a mean probability classifier ensemble, as we described in §3.1.2.2, and run against the

[18] We have 19 features and therefore $\frac{19\times18}{2} = 171$ possible combinations.

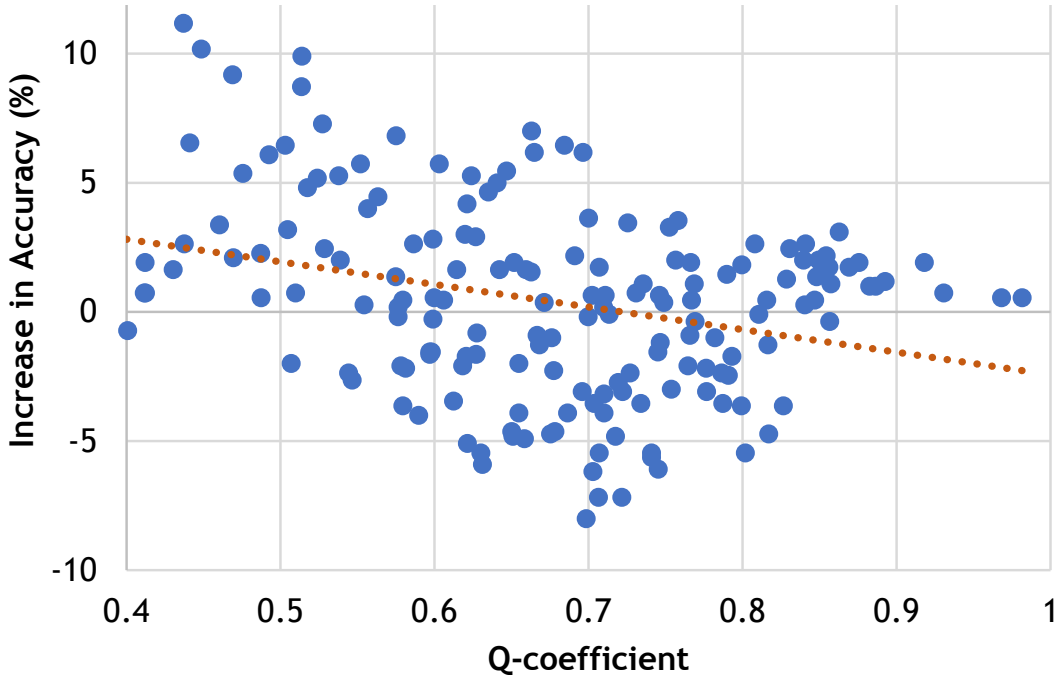

Figure 3.9: Scatterplot of the Q-coefficient vs relative increase in accuracy for all 171 feature pairs.

TOEFL11-TEST set. For each pair we also calculate the relative increase over only using the more accurate feature of the two; the relative increase is defined as:

$$Accuracy_{f_i+f_j} - max(Accuracy_{f_i}, Accuracy_{f_j})$$

This measures the net effect of combining the two: positive for improvements and negative for degradation. An alternative metric here for this could be the "Oracle" which we will introduce in Chapter 4.

The increase for each pair is compared against the Q-coefficient, and Pearson's correlation for the two variables shows a medium-sized, statistically significant negative correlation ($r = -.303$, $p = .000$). A scatterplot is shown in Figure 3.9, where we observe that almost all feature pairs with $Q < 0.5$ yielded a net increase while many pairs with $Q > 0.6$ resulted in performance degradation.

The measure is particularly useful when comparing features with similar individual accuracy to identify sets with the highest diversity. This is because diversity itself cannot be the sole criteria for feature selection; a weak feature such as character unigrams will be very diverse with respect to a strong one like POS n-grams but this does not *ipso facto* make it a good feature and we must also consider accuracy.

3.2.3 Analyzing Words and Dependencies

Grammatical dependencies have been found to be a very useful NLI feature and thought to capture a "more abstract representation of syntactic structures" (Tetreault et al., 2012; Bykh and Meurers, 2014). Accordingly, we were initially surprised to find the high correlation between dependencies and word bigrams ($Q = 0.93$). However, this relation may not be unexpected after all.

One source of supporting evidence comes from examining dependency distances. Using English data,[19] Liu (2008) reports a Mean Dependency Distance (MDD) of 2.54 with 51% of the dependencies

[19] 120k sentences averaging 21 tokens each.

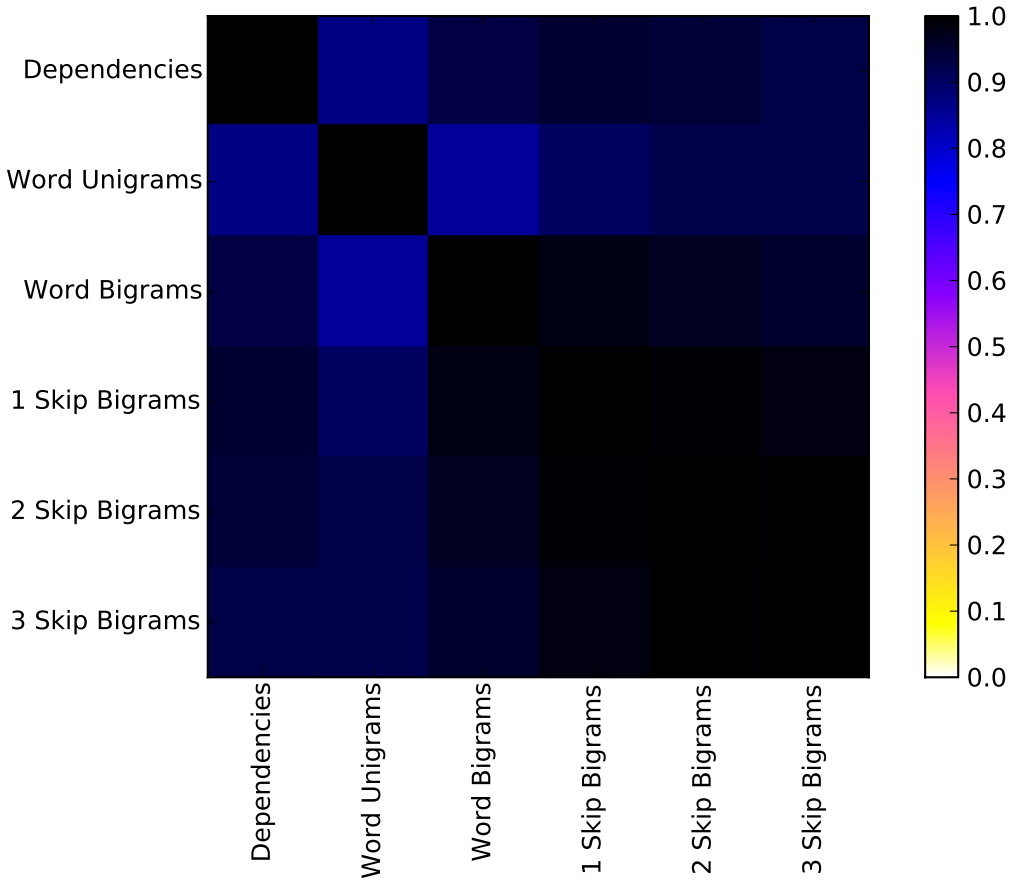

Figure 3.10: The Q-coefficient matrix for dependencies, word n-grams and skip-grams.

being adjacent and thus also captured by word bigrams. This also suggests that we can capture more of this information by considering non-adjacent tokens. We test this hypothesis by using k-skip word bigrams (Guthrie et al., 2006) as classification features, with $k = 1$–3.

The 1-skip bigrams yield an accuracy of 79.3% on the TOEFL11 test set, higher than either word bigrams or Stanford Dependencies. The 2- and 3-skip grams achieve 78.4% and 77.9%. The matrix of Q-coefficients for these features is shown in Figure 3.10, showing that the 1-skip word bigrams feature is the closest to the dependencies feature with a Q-coefficient of 0.96. It is also the closest to standard word unigrams and bigrams with Q-coefficients of 0.91 and 0.97, respectively.

These results suggest that skip-grams are a very useful feature for NLI.[20] They could also be used as a substitute for dependencies in scenarios where running a full parser may not be feasible, *e.g.* real-time data processing. Moreover, if NLI were to be applied to other languages, this feature can be a good approximation of the dependencies feature for low-resourced languages without an accurate parser. However, results may vary by language as Liu (2008) shows MDD values vary by language and possibly genre. We also note that the skip-gram feature space grows prodigiously as k increases, although this is to be expected of all n-grams.

Another related issue is whether sub-lexical character n-grams are independent of word features. Previously, Tsur and Rappoport (2007) hypothesized that these n-grams are discriminative due to writer choices "strongly influenced by the phonology of their native language". Nicolai and Kondrak (2014) also investigate the source of L1 differences in the relative frequencies of character bigrams.

[20] Hladka et al. (2013) and Henderson et al. (2013) previously used a skip-gram variant that did not include 0 skips as per Guthrie et al. (2006) and did not improve accuracy.

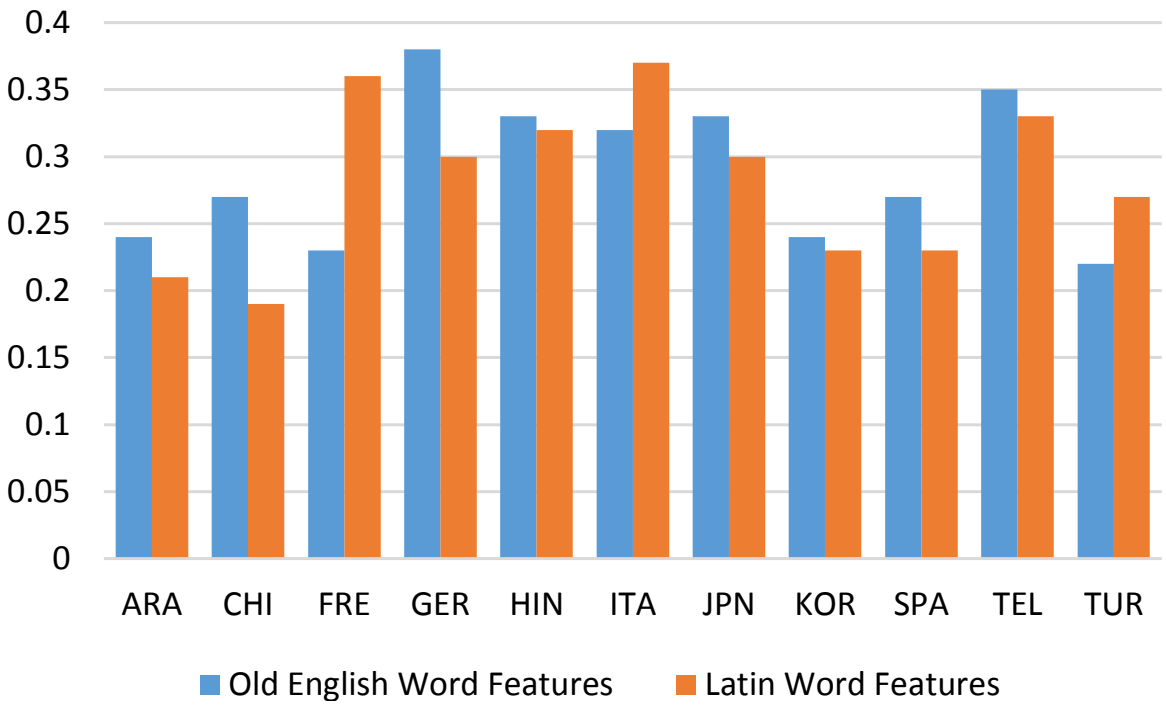

Figure 3.11: F1-scores for classifying L1 using English words with Old English or Latin origins.

They propose an algorithm to identify the most discriminative words and subsequently, the character bigrams corresponding to these words. They found that removing a small set of highly discriminative words[21] greatly degrades the accuracy of a bigram-based classifier. Based on this they conclude that bigrams capture differences in word usage and lexical transfer rather than L1 phonology. Evidence from our analysis also points to a similar pattern with the predictions of character bigrams and trigrams being strongly correlated with word and lemma unigrams.

Such lexical transfer effects have been previously noted by others (Odlin, 1989). The effects are mediated not only by cognates and word form similarities, but also semantics and meanings. We thus also examine the link between L1 and word usage.

Using the Etymological WordNet[22] database (de Melo, 2014), we extracted two lists of English words with either Old English (508 words) or Latin origins (1,310 words). These words were used as features to train two classifiers on the TOEFL11-TRAIN and TOEFL11-DEV data. The F1-scores for classification on TOEFL11-TEST set are shown in Figure 3.11. The Old English words, with their West Germanic roots, yield the best results for classifying German data. Conversely, the Latinate features achieve the best results for Italian followed by French, both languages descended from Latin. This experiment, albeit limited in scope, provides some empirical evidence suggesting that small sets of words can capture lexical transfer effects potentially mediated by L1 similarity and cognates. This relates to the lexical transfer effects that we discussed earlier in §2.4.3.

3.2.4 Conclusion

In this experiment we examined a method for measuring feature diversity in NLI and highlighted several interesting trends. We demonstrated how this analysis can be used to better understand the information captured by features and used it to examine the relationship between lexical features.

[21] Examples they provide include *opinion*, *conclude*, *however*, *France*, *Turkey*, *Italian* and *Germany*.
[22] `http://www1.icsi.berkeley.edu/%7edemelo/etymwn/`

The analysis suggested the idea, confirmed experimentally, that a variant of 1-skip bigrams can in fact be a useful feature.

3.3 Chapter Summary

In this chapter we:

- built an accurate NLI system that will serve as a key component of later chapters
- established that more linguistically fine-grained POS tagsets yield better performance
- proposed the use of function word n-gram features for NLI
- demonstrated the utility of ensemble methods for NLI
- showed that certain lexical features, *e.g.* toponyms, can exert undue influence on results
- concluded that SVM classifiers trained on frequency-based features produce the best results
- proposed a method for measuring NLI feature diversity based on Yule's Q-coefficient statistic
- highlighted several interesting trends and correlations between lexical and syntactic features

Having established a framework for NLI, in the next chapter we:

- examine aspects of NLI that are relevant to practical applications
- consider estimating the upper-bound for NLI accuracy using oracle measures
- investigate human performance for NLI using a group of experts
- assess how NLI systems perform on out-of-domain data

Chapter 4

Human Baselines, Oracles and Cross-Corpus NLI

This chapter focuses on examining aspects of NLI that are relevant to practical applications. In the first section we consider the question of determining the upper-bounds for NLI accuracy. We propose using two oracle measures to estimate this, applying them to our features and a new dataset composed of all submissions to the 2013 shared task, revealing interesting error patterns. We then consider the question of human performance for NLI, presenting the first such experiment using a group of experts. In the second section we consider how well NLI systems might perform on other data, including those from other domains and genres. We introduce and use a new dataset to conduct large-scale cross-corpus evaluation to assess this using standard features and an oracle.

Chapter Contents

Portions of the work presented in this chapter has been published as:

- Shervin Malmasi, Joel Tetreault, and Mark Dras. Oracle and Human Baselines for Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–178, Denver, Colorado, June 2015c. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-0620`

  Here the second author was the leading organizer of the 2013 Shared Task and provided all the submissions to the shared task; recruited our human experts and assisted with developing the experimental setup.

- Shervin Malmasi and Mark Dras. Large-scale Native Language Identification with Cross-Corpus Evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2015)*, pages 1403–1409, Denver, CO, USA, June 2015a. Association for Computational Linguistics. URL `http://aclweb.org/anthology/N15-1160`

4.1 Oracles and Human Performance

All of the NLI research conducted to date has relied on comparisons against simple baselines — namely the random and majority baseline measures — to evaluate if the methods surpass these baselines and can be therefore be considered "successful". An interesting question about NLI research concerns an upper-bound on the accuracy achievable for a dataset. While baselines help determine the lower bounds for performance, no attempts have been made to estimate an upper-bound. More specifically for this thesis, given a dataset, a selection of features and classifiers in an ensemble, what is the maximal performance that could be achieved by an NLI system that always picks the best candidate label? This question, not previously addressed in the context of NLI to date, is the primary focus of this section. Such a measure is an interesting and useful upper-limit baseline for researchers to consider when evaluating their work, since obtaining 100% classification accuracy may not be a reasonable or even feasible goal. In this study we investigate this issue with the aim of deriving such an upper-limit for NLI accuracy. Additionally, we use this method to to analyze the submissions from all teams in the 2013 NLI shared task.

A second goal of this section is to measure human performance for NLI, something not attempted to date. To this end we design and run a crowd-sourced experiment where expert human evaluators predict the L1 of texts from the NLI shared task and their performance is compared against an equivalent NLI system. This novel experiment provides insight into how experts would perform on this task.

4.1.1 Oracle Classifiers

One possible approach to estimating an upper-bound for classification accuracy, and one that we employ here, is the use of an "Oracle" classifier. This method has previously been used to analyze the limits of majority vote classifier combination (Kuncheva et al., 2001). An oracle is a type of multiple classifier fusion method that can be used to combine the results of an ensemble of classifiers which are all used to classify a dataset.

The oracle will assign the correct class label for an instance if at least one of the constituent classifiers in the system produces the correct label for that data point. Some example oracle results for an ensemble of three classifiers are shown in Table 4.1. The probability of correct classification of a data point by the oracle is:

$$P_{\text{Oracle}} = 1 - P(\text{All Classifiers Incorrect})$$

Oracles are usually used in comparative experiments and to gauge the performance and diversity of the classifiers chosen for an ensemble (Kuncheva, 2002; Kuncheva et al., 2003). They can help us quantify the *potential* upper limit of an ensemble's performance on the given data and how this performance varies with different ensemble configurations and combinations.

One scenario is the use of an oracle to evaluate the utility of a set of feature types. Here each classifier in the ensemble is trained on a single feature type. This is the focus of our first experiment.

Another scenario involves the combination of different learning algorithms trained on similar features, to form an ensemble in order to evaluate the potential benefits and limits of combining different classification approaches. This is the focus of our second experiment, using all of the entries from the 2013 shared task as independent systems.

		Classifier Output			
Instance	**True Label**	C_1	C_2	C_3	**Oracle**
18354.txt	ARA	TUR	ARA	ARA	Correct
15398.txt	CHI	JPN	JPN	KOR	Incorrect
22754.txt	HIN	GER	TEL	HIN	Correct
10459.txt	SPA	SPA	SPA	SPA	Correct
11567.txt	ITA	FRE	GER	SPA	Incorrect

Table 4.1: Example oracle results for an ensemble of three classifiers.

4.1.2 Ensemble Combination Methods

While the oracle chooses the correct class label if at least one constituent classifier has selected it, it is possible to generalize this idea by relating it to the combination methods of §3.1.2. The combination methods we examine are as follows.

Oracle The correct label is selected if predicted by any ensemble member, as described in §4.1.1.

Plurality Voting This is the standard combination strategy that selects the label with the highest number of votes,[1] regardless of the percentage of votes it received (Polikar, 2006), that we described in §3.1.2.1.

Accuracy@N To account for the possibility that a classifier may predict the correct label essentially by chance (with a probability determined by the random baseline) and thus exaggerate the oracle score, we propose an Accuracy@N combiner. This method is inspired by the "Precision at k" metric from Information Retrieval (Manning et al., 2008) which measures precision at fixed low levels of results (*e.g.* the top 10 results). Here, it is an extension of the Plurality vote combiner where instead of selecting the label with the highest votes, the labels are ranked by their vote counts and an instance is correctly classified if the true label is in the top N ranked candidates.[2] Another way to view it is as a more restricted version of the Oracle combiner that is limited to the top N ranked candidates in order to minimize the influence of a single classifier having chosen the correct label by chance. In this study we experiment with $N = 2$ and 3. We also note that setting $N = 1$ is equivalent to the Plurality voting method.

Mean Probability All classifiers provide probability estimates for each possible class. The estimates for each class are collected and the one with the highest mean wins (Polikar, 2006, §4.2). We previously described this method in §3.1.2.2.

Simple Combination Combines all features into a single feature space and uses a single classifier.

4.1.3 Feature Set Evaluation

Our first experiment attempts to derive the potential accuracy upper-limit of our feature set. Here we follow our previous setup described in §3.2.1, meaning that we train a single SVM classifier for

[1] This differs with a *majority* vote combiner where a label must obtain over 50% of the votes.
[2] In case of ties we choose randomly from the labels with the same number of votes.

	Accuracy (%)	
	10-fold CV	**Test Set**
Random Baseline	9.1	9.1
Shared Task Best	84.3	83.6
Oracle	95.6	95.4
Accuracy@3	92.5	92.2
Accuracy@2	88.6	88.0
Plurality Vote	78.2	77.6
Simple Combination	78.2	77.5
Mean Probability	79.4	78.7

Table 4.2: Oracle results using our feature set.

each feature type to create our classifier ensemble. We do not experiment with combining different machine learners as we did in Chapter 3, instead we focus on gauging the potential of the feature set. We employ the standard set feature types we used in Chapter 3: character/word n-grams, Part-of-Speech (POS) n-grams, function words, Context-free grammar production rules, Tree Substitution Grammar fragments and Stanford Dependencies. Detailed descriptions of these features can be found in §3.1.3.2. For features interactions and diversity, see §3.2.

We report classification accuracy under 10-fold cross-validation using the combined TOEFL11-TRAIN and TOEFL11-DEV sets. We also report accuracy on the TOEFL11-TEST set from the 2013 shared task, after training on the combined TOEFL11-TRAIN and TOEFL11-DEV sets. We compare against a random baseline and the top performer from the shared task (Jarvis et al., 2013).

The results for these experiments are shown in Table 4.2. The cross-validation and test results are very similar, with the oracle accuracy at 95%, suggesting that for each document there is in most cases at least one feature type that correctly predicts it. This drops to 88% with the Accuracy@2 combiner, but this is still substantially higher than the plurality vote and the best results from the shared task. This suggests that there is a noticeable tail of feature types dragging the plurality vote down. The results also seem to suggest that the correct label is often in the top two predictions; this is something we investigate further in the next experiment.

4.1.4 2013 Shared Task Evaluation

In the second experiment we apply our method to the submissions in the 2013 NLI Shared Task, aiming to quantify the potential upper limit for combining a range of different systems. We do not train any models here; each submission to shared task consists of predicted labels for each document in TOEFL11-DEV and is treated as an independent classifier.

The shared task data comes from the closed-training sub-task.[3] Each team was allowed to submit up to 5 different runs for each task, allowing them to experiment with different feature and parameter variations of their system. Each team's systems produce predictions using their own set of features and learning algorithms, with several of these systems using ensembles themselves.

[3]The shared task consisted of three sub-tasks. For each task, the test set was TOEFL11-TEST; only the type of training data varied by task where the other two sub-tasks allowed the use of external training data. Refer to §2.3.3.3 for more details.

	Accuracy (%)	
	Best Run	**All Runs**
Random Baseline	9.1	9.1
Shared Task Best	83.6	83.6
Oracle	97.9	99.5
Accuracy@3	95.5	95.6
Accuracy@2	92.2	92.5
Plurality Vote	84.5	84.4

Table 4.3: Oracle results on the NLI 2013 shared task systems. The Best Run ensemble includes each team's best system, for a total of 29 systems. The All Runs ensemble consists of all 116 submission from all teams.

In total, 116 runs were submitted by 29 teams, with the winning entry achieving the highest accuracy of 83.6% on the TOEFL11-TEST set. The data for this experiment was prepared by aggregating all submitted runs. Additionally, we have also made this data, along with the entries for the other open training sub-tasks, available for use by other researchers. The predictions can be downloaded[4] in CSV format. This data, known as the NLI2013 SUBMISSIONS dataset, can assist with research in NLI as well as more fundamental machine learning and ensemble research.

The ensembles here are constructed by combining the submitted entries, each of which is treated as a single system. We experiment under two conditions: using only each team's best run (29 systems in total) and using all 116 runs together. Results are compared against the random baseline and winning entry in the competition.

Table 4.3 shows the results for this experiment. The oracle results here are higher than the previous experiment, which is not unexpected given the much larger number of predictions per document. Results for the other combiners are also higher here.

The Accuracy@2 results are 92% in both conditions, much higher than the winning entry's 83.6%. Results from the Accuracy@2 combiner, both here and in the previous experiment, show that a great majority of the texts are close to being correctly classified: this value is significantly higher than the plurality combiner[5] and not much lower than the oracle itself. This shows that the correct label receives a significant portion of the votes, and when not the winning label, it is often the runner-up.[6]

One implication of this concerns practical applications of NLI, *e.g.* in a manual analysis, where it may be worthwhile for researchers to also consider the runner-up label in their evaluation. When using NLI in critical applications in an operational (*i.e.* real-world) setting, it may be more prudent to employ a semi-automatic, computer-aided classification system (Larkey and Croft, 1996) instead of a fully-automatic classifier.

This knowledge could also be used to increase NLI accuracy by aiming to develop more sophisticated classifiers that can take into account the top N labels in their decision making, similar to discriminative reranking methods applied in statistical parsing (Charniak and Johnson, 2005).

We also look more closely at the relationship between the top two label values. Using the Accuracy@2 combiner, we isolate the cases where the actual label was the runner up and extract the most frequent pairs of such top 2 labels, presented in Table 4.4. We see that a quarter of these errors are

[4] `http://web.science.mq.edu.au/~smalmasi/data/NLI2013-SUBMISSIONS.zip`
[5] Which is itself equivalent to an Accuracy@1 combiner.
[6] In approx. 8% of the cases here, to be more precise.

confusion between Hindi and Telugu. Chinese, Japanese and Korean are also often confused by the shared task entries.

We also examine the confusion matrices for the plurality, Accuracy@2 and oracle combiners,[7] shown in Figure 4.1. They demonstrate that Hindi–Telugu is the most commonly confused pair, as we discussed in §3.1.5.2, and confirm the directionality of the confusion: more Telugu texts are misclassified as Hindi than vice versa. The oracle confusion matrix highlights the few cases where none of the systems could correctly predict the label.

4.1.5 Human NLI Performance

While most NLI work has focused on improving system performance, to our knowledge there has not been any corresponding study which looks at human performance for this task. To give our preceding results more context, as well as the results of the field, we ran an exploratory study to determine how accurate humans are for this task.

4.1.5.1 Experiment Design

Our initial idea was to use the Amazon Mechanical Turk to collect crowdsourced judgments. However, unlike simpler NLP tasks, *e.g.* sentiment analysis and word sense disambiguation, which can be effectively annotated by untrained Turkers (Snow et al., 2008), NLI requires raters with knowledge and exposure to writers with different L1s. These constitute a very specific skill set that will be hard to obtain using untrained Turkers.

Optimally, one would use a set of ESL teachers and researchers who have experience in working with ESL writers from all of the 11 L1s, though such people are rarity. As a reasonable compromise, we chose 10 professors and researchers who have varied linguistic backgrounds, speak multiple languages, and have had exposure with the particular L1s, either as a speaker or through working with ESL students. We also constrained the task from 11 L1s to 5 (Arabic, Chinese, German, Hindi, and Spanish) as we believed that 11 L1s would be too much of an overload on the judges. The 5 L1s selected all belong to separate language families.

The experiment consisted of rating 30 essays from TOEFL11-TEST. These essays were split into two conditions: 15 of them we selected from the subset which most Shared Task systems could predict correctly (easy), and the remaining 15 were essays which the most Shared Task systems had difficulty predicting (hard). The L1s were distributed evenly over the essays and both easy and hard conditions (3 "easy" and 3 "hard" essays per L1). All raters completed the task within an hour.

[7] Where the Accuracy@2 and oracle combiners could not predict the correct label the plurality vote was used instead.

Confused Pair	Percent	Cumulative Percent
HIN–TEL	15.9	15.9
TEL–HIN	10.2	26.1
CHI–KOR	6.8	33.0
JPN–KOR	6.8	39.8
KOR–TUR	4.5	44.3

Table 4.4: The most commonly predicted top 2 class label pairs where the runner-up is the true label.

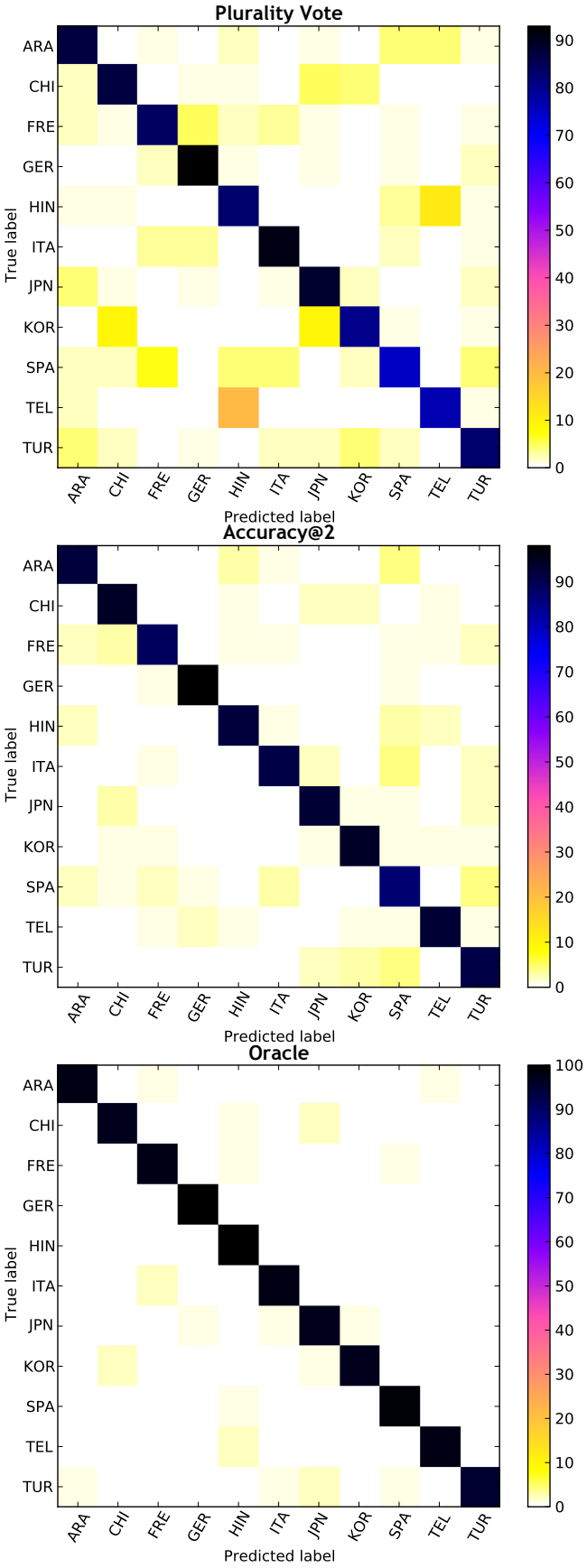

Figure 4.1: Confusion matrices for the plurality (top), Accuracy@2 (middle) and oracle (bottom) combiners.

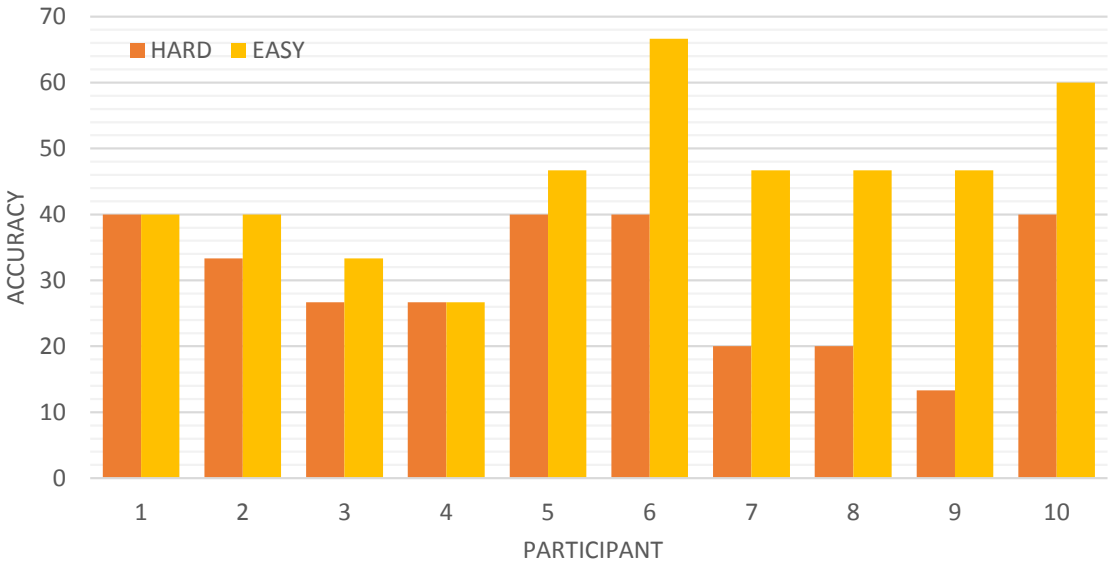

Figure 4.2: Prediction accuracy for each of our 10 participants under both easy and hard conditions.

	Accuracy (%)		
	Easy	**Hard**	**All**
Random Baseline	20.0	20.0	20.0
NLI Plurality Vote	100.0	33.3	66.7
NLI Mean Probability	**100.0**	**46.7**	**73.3**
Top Human	66.7	40.0	53.3
Human Plurality Vote	73.3	40.0	56.7

Table 4.5: Comparing human participant performance against an NLI system on 30 selected texts.

4.1.5.2 Results

Figure 4.2 shows the accuracy for each of the 10 raters in this pilot study. The top rater accurately identified the L1 for 16 out of 30 texts (53.3%), with the lowest raters at 30.0% overall and an average of 37.3%. All raters did at least as well on the "easy" cases than on the "hard", although usually better. A more detailed version of the results with individual accuracy for all 10 raters across all texts is shown in Figure 4.3.

A paired-samples t-test was conducted to compare human accuracy in the easy and hard conditions. A significant difference was found for easy (M=45.33, SD=11.67) and hard (M=30, SD=10.06), t(9)=−3.851, p = .004.

Next, we compared human accuracy with our NLI system, which we re-trained using only the five selected L1s. Results are shown in Table 4.5. All ensembles outperform human raters and a plurality vote composed of the human raters. Interestingly, the human plurality vote was only 3% higher than the top human score, suggesting that the raters tended to get the same essays correct.

We also note that some L1s received more correct predictions than others,[8] but the difference is not statistically significant.[9] A confusion matrix generated from the plurality vote of the human experts, shown in Figure 4.4, also reflects this pattern. It shows that Chinese is the most correctly predicted class (followed by Spanish) while German is the most confused. This can be compared

[8] CHI: 50%, SPA: 46.7%, HIN: 33.3%, GER: 31.7%, ARA: 26.7%
[9] Our sample size is too small, but this is still suggestive.

	Gold Label	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	Rater 10	Correct Shared Task Predictions
Text 1	SPA	GER	SPA	CHI	SPA	ARA	SPA	ARA	CHI	ARA		29
Text 2	ARA	SPA	CHI	ARA	HIN	ARA	ARA	SPA	CHI	ARA		29
Text 3	ARA	HIN	CHI	CHI	HIN	ARA	ARA	HIN	CHI	ARA		29
Text 4	GER	HIN	HIN	HIN	CHI	ARA	ARA	HIN	GER	GER		29
Text 5	CHI	HIN	CHI	GER	GER	HIN	GER	GER	CHI	CHI		29
Text 6	CHI	CHI	CHI	CHI	CHI	CHI	CHI	ARA	CHI	CHI		29
Text 7	CHI	CHI	ARA	SPA	GER	CHI	HIN	CHI	ARA	GER		29
Text 8	GER	GER	ARA	ARA	ARA	ARA	GER	HIN	HIN	HIN		29
Text 9	GER	HIN	GER	HIN	GER	ARA	CHI	GER	GER	GER		29
Text 10	ARA	HIN	CHI	CHI	CHI	ARA	ARA	CHI	ARA	ARA		29
Text 11	SPA	HIN	SPA	SPA	SPA	SPA	SPA	GER	CHI	SPA		29
Text 12	HIN	HIN	GER	GER	HIN	GER	GER	HIN	HIN	SPA		28
Text 13	SPA	SPA	SPA	GER	ARA	GER	SPA	HIN	GER	ARA	CHI	28
Text 14	HIN	ARA	GER	GER	GER	SPA	HIN	GER	SPA	GER	SPA	28
Text 15	HIN	HIN	GER	HIN	ARA	GER	HIN	HIN	GER	HIN	HIN	28
Text 16	GER	HIN	ARA	SPA	SPA	ARA	SPA	GER	ARA	CHI	ARA	5
Text 17	ARA	GER	GER	SPA	SPA	GER	ARA	GER	SPA	HIN	GER	4
Text 18	GER	GER	GER	HIN	HIN	GER	GER	GER	CHI	HIN	HIN	4
Text 19	CHI	ARA	HIN	SPA	CHI	GER	HIN	GER	CHI	ARA	ARA	3
Text 20	HIN	HIN	HIN	SPA	CHI	CHI	HIN	GER	CHI	SPA	HIN	2
Text 21	SPA	HIN	SPA	CHI	SPA	SPA	CHI	CHI	GER	CHI	CHI	2
Text 22	SPA	SPA	ARA	CHI	SPA	SPA	SPA	GER	HIN	SPA	SPA	1
Text 23	HIN	ARA	HIN	HIN	CHI	GER	CHI	SPA	SPA	CHI	GER	1
Text 24	HIN	HIN	CHI	CHI	SPA	ARA	CHI	ARA	CHI	ARA	HIN	1
Text 25	SPA	HIN	ARA	CHI	SPA	SPA	GER	SPA	SPA	ARA	SPA	1
Text 26	GER	GER	HIN	GER	ARA	CHI	ARA	HIN	CHI	SPA	CHI	1
Text 27	CHI	HIN	SPA	CHI	HIN	SPA	CHI	ARA	CHI	CHI	CHI	1
Text 28	CHI	CHI	CHI	CHI	ARA	CHI	CHI	ARA	HIN	ARA	CHI	0
Text 29	ARA	GER	GER	SPA	GER	GER	GER	GER	HIN	GER	HIN	0
Text 30	ARA	SPA	HIN	GER	HIN	ARA	HIN	HIN	SPA	SPA	HIN	0

Figure 4.3: Individual results for the 10 raters across all texts. Correct predictions are highlighted in green. The last column indicates how many shared task entries correctly predicted that text.

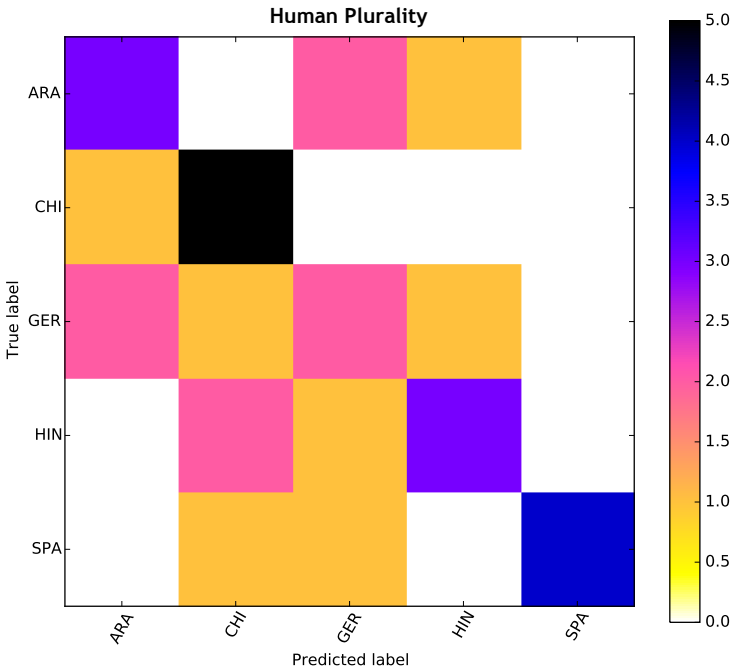

Figure 4.4: Confusion matrix based on a plurality vote using the human expert predictions on all 30 learner essays.

against the confusion matrix of the NLI system using the mean probability ensemble, as shown in Figure 4.5.

Some participants noted that while they had familiarity with L1 Spanish and Chinese non-native writing, they did not have much exposure to the other L1s, possibly due to international student cohorts.

Our belief, based on these pilot results, is that as the number of classes increases, the machine learning systems will outpace the human raters by an increasingly wide margin. It should also be noted that we purposefully selected disparate L1s to make easier for the human raters. As there are several other L1s in the TOEFL11 that are in the Romance family class, and others where it is less likely for raters to have seen student essays (such as Telugu), including those will also likely affect human performance.

4.1.6 Discussion

We presented a novel analysis for predicting the "potential" upper limit of NLI accuracy on a dataset. This upper limit can vary depending on which components — feature types and algorithms — are used in the system. Alongside other baselines, oracle performance can assist in interpreting the relative performance of an NLI system.[10]

Using the entries submitted by the participants of the 2013 NLI shared task we conducted the largest analysis of NLI systems to date. An analysis using our Accuracy@N oracle showed that a significant portion of the errors being committed involve the confusion of the top two labels. Future research that focuses on improving classification for these frequently confused class pairs can lead to very large improvements in overall results.

[10] *e.g.* an NLI system with 70% accuracy against an Oracle baseline of 80% is relatively better compared to one with 74% accuracy against an Oracle baseline of 93%.

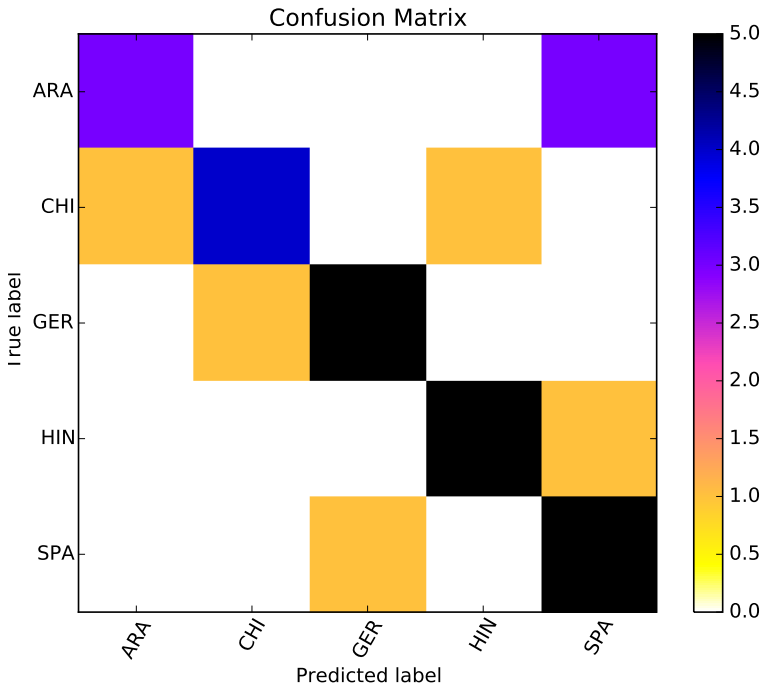

Figure 4.5: Confusion matrix based on the NLI systems predictions for all 30 learner essays.

We also ran the first study of human performance for NLI, with results suggesting a committee of 10 raters could not outperform our NLI system on a simplified version of the task. Our experience in designing the experiment also highlighted the difficulties in finding experts suited for this task.

There are also additional issues to consider for future experiments using human judges. An important question to raise is whether the professors and researchers used in this study are the best judges for this task. Would experienced ESL teachers perform better? Were the experts chosen for this study a good approximation for such ESL teachers? It could be argued that speaking a second language and/or interacting with a small number of students who are non-native speakers is not the same as reading thousands of learner essays over a long period and noticing characteristic errors. From this point of view our systems received more "training" than the human experts here. However, we would still contend that finding such experts with significant experience across a large number of L1 is not an easy undertaking in itself.

A useful application of the oracle method is to isolate the subset of wholly misclassified texts for further investigation and error analysis. This segregated data can then be independently studied to better understand the aspects that make it hard to classify them correctly. This can also be used to guide feature engineering practices in order to develop features that can distinguish these challenging texts. In practice, this type of oracle measure can be used to guide the process of choosing the pool of classifiers that form an ensemble.

We also note that these oracle figures would be produced by an optimal system that always makes the correct decision using this pool of classifiers. While these oracle results could be interpreted as potentially attainable, this may not be feasible and practical limits could be substantially lower.

4.2 Large-Scale Cross-Corpus Evaluation

The work we have examined so far has focused mostly on a single corpus, using cross-validation and test sets for evaluation. Alternatively, cross-corpus evaluation involves training one one corpus and testing on a completely different corpus. This approach can help assess the fundamental issue of how much the trained model generalizes beyond the particular characteristics of the training data.

Furthermore, deriving SLA hypotheses from a single corpus may not be entirely useful for SLA research. Many variables like genre and topic are constant within a corpus, restricting the validity of such cross-validation studies to those dimensions.

An alternative, potentially more helpful approach, is to identify transfer features that reliably distinguish an L1 across multiple corpora of differing genres and domains. A cross-corpus methodology may be a more promising avenue to finding features that generalize to diverse text sources, but requires additional large corpora. It is also a more realistic approach,[11] and one we pursue in this work.

Accordingly, the aims of the present experiment are to: (1) test a large new corpus suitable for NLI; (2) perform within-corpus evaluation with a comparative analysis against equivalent corpora; (3) perform cross-corpus evaluation to determine the efficiency of corpus independent features; and (4) examine the possibility of using lexical features in a cross-corpus scenario.

4.2.1 Related Work

Cross-corpus studies have been conducted for various data-driven NLP tasks and applications, including parsing (Gildea, 2001), WSD (Escudero et al., 2000), NER (Nothman et al., 2009) and biomedical text mining (Airola et al., 2008). While most such experiments show a drop in performance, the effect varies widely across tasks, making it hard to predict the expected drop for NLI. We aim to address this question using large training and testing data.

The use of cross-corpus evaluation in NLI has been limited thus far, likely due to the paucity of suitable data; very few NLI studies have been conducted using distinct corpora for training and testing.

Brooke and Hirst (2011) created the Lang-8 corpus by extracting a large collection of online journal entries from the Lang-8 website,[12] an online platform designed to assist learners improve their language skills by writing such entries. They conduct a cross-corpus experiment using 7 L1 classes with 200 texts per class per corpus. Comparing against a 14.3% random baseline, they achieve 26.7% accuracy for training on the ICLE corpus (§2.3.1.2) and testing on the Lang-8 data and 46.1% for the reverse setup. They use a combination of several standard features,[13] including lexical ones, with word n-grams being the most accurate. They note that performance is much lower than cross-validation.

Bykh and Meurers (2014) explore the use of lexicalized and non-lexicalized phrase structure rules for NLI. They build their own dataset for cross-corpus evaluation by amalgamating data from five different corpora. This dataset consists of 5,843 texts across the same 11 classes as the TOEFL11, although their data is unbalanced.[14] Despite their data not being topic balanced, they use lexical features and report cross-corpus accuracies of 28–43% against a majority baseline of 18%.

[11] *i.e.* when applied in practical scenarios, real world data being evaluated is unlikely to be identical in nature to the training corpora used here.
[12] `http://lang-8.com/`
[13] Function words, POS n-grams, Character n-grams and word n-grams.
[14] *e.g.* they have over 1000 L1 Chinese texts and only 349 L1 Turkish texts.

EFCamDat	**TOEFL11**
850 texts per L1	1,100 texts per L1
Arabic	Arabic
Chinese	Chinese
French	French
German	German
Italian	Italian
Japanese	Japanese
Korean	Korean
Spanish	Spanish
Turkish	Turkish
Portuguese	Hindi
Russian	Telugu

Table 4.6: The 11 L1 classes extracted from the EFCAMDAT corpus, compared to the TOEFL11 corpus. The first 9 classes are common between both.

In sum, previous work has relied mostly on small and unbalanced datasets in cross-corpus evaluation. In this section we aim to improve on this by using large and balanced corpora for both training and testing.

Furthermore, although these studies have used lexical features, they have not examined (*e.g.* via feature analysis) the specific words which are informative in a cross-corpus context in order to verify that no types of bias, topic or otherwise, exist. Tetreault et al. (2012) justify their use of lexical features by noting that their corpora "are composed of sufficiently different topics that there should not be significant overlap". However, as we discovered in our features analysis of the shared task data in §3.1.5.1, other categories of non-topical words, such as the toponyms, can exert undue influence on classification results. This is an issue which we will investigate in this experiment.

4.2.2 EFCamDat: A new corpus for NLI

The EF Cambridge Open Language Database (EFCAMDAT) is an English L2 corpus that was released recently (Geertzen et al., 2013). It is composed of texts submitted to *Englishtown*, an online school used by thousands of learners daily.

This corpus is notable for its size, containing some 550k texts from numerous nationalities, making it an ideal candidate for NLI research. While the TOEFL11 is made of argumentative essays, EFCAMDAT has a much wider range of genres including writing emails, descriptions, letters, reviews, instructions and more.

In this section we present an application of NLI to this new data. As some of the texts can be short, we use the methodology of Brooke and Hirst (2011) to concatenate and create texts with at least 300 tokens, much like the TOEFL11.

From the data we choose 850 texts from each of the top 11 nationalities.[15] This subset of EFCAMDAT thus consists of 9,350 documents totalling approximately 3.2m tokens. This is an average of 337 tokens per text, close to the 348 tokens per text in TOEFL11.

[15]This was the largest number for which we could get documents from all 11 classes. A higher amount would have resulted in an imbalanced dataset.

This also provides us with the same number of classes as the TOEFL11, as shown in Table 4.6, facilitating direct performance comparisons. The table also indicates the 9 classes common to both corpora. This subset of common classes enables us to perform large-scale cross-corpus validation experiments that have not been possible until now.

4.2.3 Methodology

We use the same classification approach we described in §3.2.1, *viz.*, a linear SVM is used trained for each feature type using relative frequency values. We also combine features with a mean probability ensemble classifier which we described in §3.1.2.2. We compare results against a random baseline as well as an oracle.

We use the EFCAMDAT data, described above in §4.2.2, as well as the the TOEFL11 data, which is a combination of the TOEFL11-TRAIN and TOEFL11-DEV sets here.

We use a set of well-tested and basic features, as the focus of our experiment here is cross-corpus evaluation and not maximizing accuracy. Due to the potential topic bias issues described in §2.3.1.1, we avoid using lexical features as EFCAMDAT is not topic balanced. We extract the following topic-independent feature types:

Function words are extracted as described in §2.3.2.5. We also apply function word bigrams as described in §3.1.3.4.

Context-free Grammar Production Rules are extracted after parsing each sentence, as described in §2.3.2.7 and §3.1.3.2.

Part-of-Speech (POS) n-grams of size 1–3 are also extracted as features, as we described in §2.3.2.6 and §3.1.3.2. We use the Penn Treebank tagset here.

4.2.4 Within-Corpus Evaluation

Our first experiment applies 10-fold cross-validation within the EFCAMDAT corpus to assess feature efficacy. The results are shown in the first column of Table 4.7.

All features perform substantially higher than the 9% chance baseline. POS trigrams are the best single feature (53%), suggesting there exist significant interclass syntactic differences. Combining all the features using an ensemble yields the best accuracy of 65% against an upper-bound of 87% set by the oracle.

We also compare these within-corpus results to those from the TOEFL11. As shown in Figure 4.6, we find that feature performance is nearly identical across corpora. We also note that the oracle performance on the TOEFL11 data using the same feature set is almost identical (86.2%), something which we will revisit in §5.9.

Figure 4.7 shows the confusion matrix for the results obtained with the ensemble. German is the most correctly classified L1, while the highest confusion is between Japanese–Korean, followed by Spanish–Portuguese and French–Italian. This is not surprising given their syntactic similarity as well as being typologically related in case of the latter two.

These trends are very similar to those we observed in the TOEFL11 data, as previously shown in the confusion matrix in Figure 3.5. Another interesting observation here is that German is the most

Classification Feature	EfCamDat 10-fold CV	Train EfCamDat Test Toefl11	Train Toefl11 Test EfCamDat
Random Baseline	9.09	11.11	11.11
Oracle	86.84	64.92	62.43
Function Word unigrams	52.01	27.14	21.77
Function Word bigrams	47.92	29.21	22.63
Production Rules	49.12	30.73	23.91
Part-of-Speech unigrams	33.21	23.42	16.71
Part-of-Speech bigrams	50.43	31.02	23.09
Part-of-Speech trigrams	53.05	32.38	25.55
Ensemble (All features)	**64.95**	**33.45**	**28.42**
Word unigrams	–	41.82	42.48

Table 4.7: Classification accuracy (%) for our within- and cross-corpus experiments. The three result columns are: (1) cross-validation within the EfCamDat, (2) training on the EfCamDat and testing on Toefl11 and (3) training on Toefl11 and testing on EfCamDat.

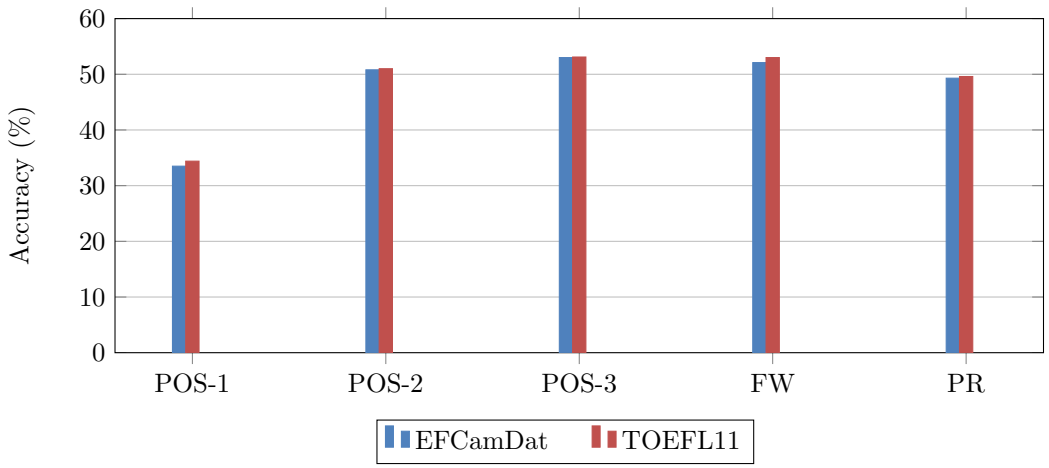

Figure 4.6: Comparing feature performance on the EfCamDat corpus against the Toefl11 corpus. POS-1/2/3: POS uni/bi/trigrams, FW: Function Words, PR: CFG Productions

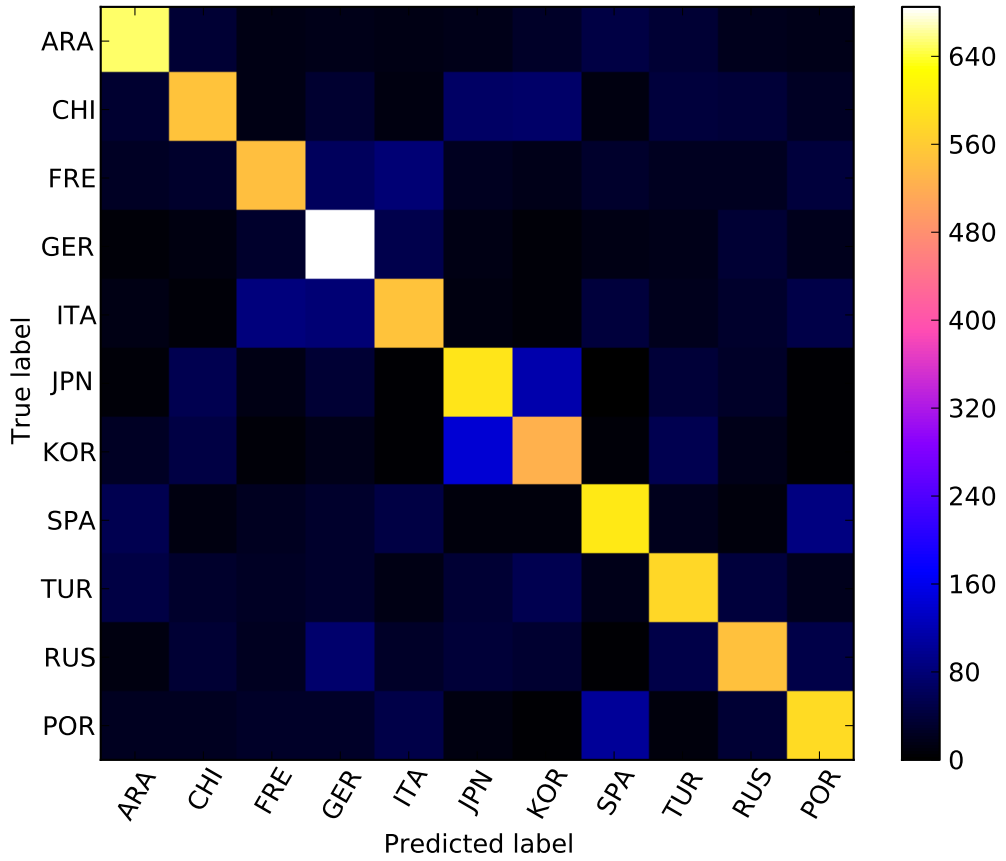

Figure 4.7: EFCAMDAT 11-class confusion matrix.

correctly predicted class in both datasets. One possible explanation is that German is typologically the closest language to the target L2 – English – across both datasets, resulting in linguistic patterns that are possibly closer to native English writing than the other L1s represented. Pursuing this line of inquiry could be done through additional experiments, for example, by adding a class of native English data to the experiment.[16] If our intuition is valid, then German L1 texts would have the greatest amount of confusion with the native texts.

4.2.5 Large-Scale Cross-Corpus Evaluation

Our second experiment tests the cross-corpus efficacy of the features by training on EFCAMDAT and testing on the TOEFL11,[17] and *vice versa*. As the corpus texts are from different genres, this approach enables us to test the cross-corpus and cross-genre generalizability of our features.

Having established the oracle accuracy for cross-validation, we will also compare this against the cross-corpus oracle performance.

Results are shown in the second and third columns of Table 4.7. While lower than the cross-validation results which were on 11 classes *vs* 9 here, the results are still far greater than the baseline. The accuracy for training on EFCAMDAT and testing on TOEFL11 is higher (33.45%) than the other way around (28.42%), even though TOEFL11 is the slightly larger corpus. This is possibly because EFCAMDAT has numerous genres while TOEFL11 does not. The cross-corpus oracle is also over 20% lower than the cross-validation results, despite an increase in the random baseline, showing that some features are not portable across corpora. Training on TOEFL11 also yields a lower oracle.

[16] This native data would need to be controlled for topic and genre in order to perform a comparable analysis.
[17] The 9 common classes discussed in §4.2.2 are used.

Arabic	German	Japanese
Saudi	Germany	Japan
Arabia	Berlin	Tokyo
Arabic	Hamburg	Osaka
Mohammed	Frankfurt	Nagoya
Ali	Munich	Yen

Table 4.8: Selected items from the top 15 most discriminative words for Arabic/German/Japanese.

Although a performance drop was expected due to the big genre differences, results suggest the presence of some corpus-independent features that capture cross-linguistic influence. However, they also suggest that a large portion of the features helpful for NLI are genre-dependent.

4.2.6 Lexical Feature Analysis

Previously, word n-grams have been applied in small-scale cross-corpus studies and found to be the best feature, as we discussed earlier in §4.2.1. Word n-grams have been previously used in NLI and are believed to capture lexical transfer effects which have been previously noted by researchers and linguists (Odlin, 1989). The effects are mediated not only by cognates and word form similarities, but also semantics and meanings. Other NLI studies have also provided empirical evidence for this hypothesis and agree with our results in §3.2.3.

However, issues stemming from topic bias[18] — which we described in §2.3.1.1 — have also limited their use in NLI. Although their use has been justified in cross-corpus scenarios due to the lower risk of topic overlap across corpora, we earlier noted that issues with using lexical features extend beyond topics to words such as toponyms. In an attempt to investigate this, we applied word unigrams to our cross-corpus experiment. We achieved an accuracy of 41.8% for training on the EFCAMDAT and testing on TOEFL11 and 42.5% for the reverse setting, as shown in the last row of Table 4.7. These are the best results in this setup.

To check for any topic-bias effects, we inspected the most discriminative features for each L1 class. This was done by ranking the features according to the weights assigned by the SVM model. In this manner, SVMs have been successfully applied in data mining and knowledge discovery tasks such as identifying discriminant cancer genes (Guyon et al., 2002). The specifics of this feature ranking method are specified in §6.3.

This analysis revealed that the top features were mostly cultural and geographic references related to the author's country. Table 4.8 contains words selected from the top 15 most discriminative features found in the cross-corpus experiment for three L1s. We observe that most of these are toponyms or culture-specific terms such as names and currencies. These results reveal another potential issue with using lexical features. Although this isn't topic-bias, the features do not represent genuine linguistic differences or lexical transfer effects between L1 groups. This type of analysis was not undertaken by the previous word we outlined in §4.2.1. In practical scenarios such as described in §1.1, this could also make NLI systems vulnerable to content-based manipulation. The exclusion of proper nouns or certain named entities from the extracted features is one potential way to combat this.

[18] Due to correlations between text topics and L1 classes.

4.2.7 Discussion

In this section we presented the first application of NLI to one of the largest and newest publicly available L2 English learner corpora. Cross-validation experiments mirrored the performance of other corpora and demonstrated its utility for the task. We believe this will motivate future work by equipping researchers with a large-scale corpus that is highly suitable for NLI.

Next, results from the largest cross-corpus NLI evaluation to date were presented, providing evidence for the presence of transfer features that generalize across learners, corpora, topics and genres. However, the fact that the cross-corpus accuracy is lower than within-corpus cross-validation highlights that a large portion of the features are highly corpus-specific. This suggests that NLI models are not entirely portable across corpora.

This has implications for applications of NLI in the real world and warrants further analysis. It is a critical issue for NLI to resolve whether (and if so, why) the features are corpus specific – and to what extent these are corpus selection effects or the consequences of our approach to testing writing skills using open-ended prompts that result in idiosyncratic content (such as mentioning stories discussing home or cultural experiences).

Practical applications of NLI to forensic linguistics or SLA must be robust to input from numerous sources and their associated variations, and this finding highlights the need for a cross-corpus approach.

We also employed the oracle classifier in a cross-corpus setting and showed that this trend was also reflected by its substantially lower accuracy. Based on this we believe that the oracle can be a very useful method for evaluating cross-corpus experiments.

Finally, we conducted a cross-corpus lexical feature analysis and demonstrated that the previous arguments for using lexical features in such experiments due to absence of topic overlap may not be entirely valid.

A shortcoming here is that we did not balance texts by proficiency to match the TOEFL11. We expect that a more even sampling of proficiency or using proficiency-segregated models will yield higher accuracy and features more representative of students at each proficiency level.

4.3 Chapter Summary

In this chapter we:

- proposed two oracles to estimate upper bounds for NLI accuracy
- released a new dataset composed of all the submissions to the 2013 NLI Shared Task
- used the oracles, this data and our feature to analyze NLI performance
- established that a substantial portion of errors involve the confusion of the top two labels
- conducted the first study of human performance for the task of NLI using 10 experts
- showed that experts could not outperform an NLI system on a simplified version of the task
- tested a large new corpus for NLI, showing that performance is similar to the TOEFL11 data
- performed a large-scale cross-corpus experiment using the TOEFL11 and EFCAMDAT corpora
- analyzed the results and showed that NLI features may not generalize well across genres
- explored the lexical features that were highly informative in a cross-corpus setting
- verified that lexical features like names and toponyms can exert undue influence on NLI results

Following this, in the next chapter we:

- explore the cross-lingual applicability of NLI by extending it to other L2 languages
- evaluate if previous results, *e.g.* large tagsets, feature diversity & oracles, hold across languages
- examine the potential for distinguishing native and non-native writing in multiple languages

Chapter 5

Multilingual NLI: Exploring Other L2s Beyond English

Our work thus far has looked at NLI on L2 English data and evaluated various aspects of what makes it effective. In this chapter we present the first large-scale extension of NLI to other L2s, using data from six new languages. A cornerstone of this chapter is the identification of suitable data for use here and in future work: we have collected corpora used here from a wide range of sources. We use standard sets of features to systematically compare results across the set of L2s and a wide range of L1s. We also evaluate if other result patterns, *e.g.* effects of POS tagset size, oracle performance and feature diversity, also hold across languages. We also employ the NLI setup in an experiment aiming to distinguish the writings of native speakers from non-natives, using data from three languages.

Chapter Contents

The work presented in this chapter has been published as:

- Shervin Malmasi and Mark Dras. Multilingual Native Language Identification. *Natural Language Engineering*, FirstView:1–53, December 2015d. ISSN 1469-8110. doi: 10.1017/S1351324915000406
- Maolin Wang, Shervin Malmasi, and Mingxuan Huang. The Jinan Chinese Learner Corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 118–123, Denver, Colorado, June 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-0614`

5.1 Motivation

While it has attracted significant attention from researchers, almost all of the NLI research to date, which we reviewed in §2.3.3, has focused exclusively on English L2 data. In fact, most work in SLA, and NLP for that matter has dealt with English. This is largely due to the fact that since World War II, the world has witnessed the ascendancy of English as its *lingua franca*. While English is the native language of over 400 million people in the U.S., U.K. and the Commonwealth, there are also over a billion people who speak English as their second or foreign language (Guo and Beckett, 2007). This has created a global environment where learning multiple languages is not exceptional and this has fuelled the growing research into second language acquisition.

However, while English is one of the most widely spoken languages in the world there are still a sizeable number of jobs and activities in parts of the world where the acquisition of a language other than English is a necessity.

One such example is Finland, where due to the predicted labour shortage, the government has adopted policies encouraging economic and work-related migration (Ministry of Labour, 2006), with an emphasis on the role of the education system. Aiding new immigrants to learn the Finnish language has been a key pillar of this policy particularly, as learning the language of the host nation has been found to be an important factor for social integration and assimilation, especially in Finland (Nieminen, 2009). This, in turn, has motivated research in studying the acquisition of Finnish to identify the most challenging aspects of the process.[1]

Another such example is that of Chinese. Interest in learning Chinese is rapidly growing,[2] leading to increased research in Teaching Chinese as a Foreign Language (TCFL) and the development of related resources such as learner corpora (Chen et al., 2010). This booming growth in Chinese language learning (Rose and Carson, 2014; Zhao and Huang, 2010), related to the dramatic globalization of the past few decades and a shift in the global language order (Tsung and Cruickshank, 2011), has brought with it learners from diverse backgrounds. Consequently, a key challenge here is the development of appropriate resources — language learning tools, assessments and pedagogical materials — driven by language technology, applied linguistics and SLA research (Tsung and Cruickshank, 2011).

Yet another case is the teaching of Arabic as a foreign language which has experienced unparalleled growth in the past two decades. For a long time the teaching of Arabic was not considered a priority, but this view has now changed. Arabic is now perceived as a critical and strategically useful language (Ryding, 2013), with enrolments rising rapidly and already at an all time high (Wahba et al., 2013).

These trends of focusing on other languages is also reflected in the NLP community, evidenced by the continuously increasing research focus on tools and resources for languages like Arabic (Habash, 2010) and Chinese (Wong et al., 2009).

Given the increasing research focusing on other L2s, we believe that there is a need to apply NLI to other languages, not only to gauge their applicability but also to aid in teaching research for other emerging languages. This need is partially driven by the increasing number of learners of other languages, as described above.

An important question that arises here is how linguistic and typological differences with English may affect NLI performance. The six languages investigated in this chapter vary significantly with respect not only to English, but also amongst themselves, in various linguistic subsystems; these dif-

[1] For example, the recent study by Siitonen (2014).

[2] `http://www.china.usc.edu/chinese-language-study-rising-fast`

ferences are detailed in §5.2. In this regard the current study aims to assess whether these differences significantly impact NLI performance for different L2s.

5.1.1 Goals and Objectives

This chapter has several aims related to the work from the previous chapters as well as other aspects of NLI. The overarching goal here is to experiment with the extension of NLI to languages other than English. One objective is to investigate the efficacy of the features that have been common to almost all NLI approaches to date, including our work thus far, on several languages which are significantly different from English. Answering this question requires the identification of the relevant non-English learner corpora. This data is then used in our first two experiments to assess whether NLI techniques and features work across a diverse set of languages. Having identified the required corpora, our next objective here is to use cross-lingual evidence to investigate other core issues in NLI.

While NLI research has investigated the characteristics that distinguish L1 groups, this has not been wholly extended to automatically discriminating native and non-native texts. This extension, using appropriate native control data, is another aim of the work presented in this chapter.

Another issue that arises in this type of multilingual research is the use of multiple part-of-speech (POS) tagsets developed for different languages. Differences in the granularity of the tags mean that they are often not directly comparable. Notwithstanding comparability, in §3.1 we also concluded that a more linguistically fine-grained POS tagset produced better results for English NLI. Both of these issues are investigated by one of our experiments where we compare the performance of different tagsets and also convert the tags for each language to a more general and common tagset in order to make the results more directly comparable across the languages.

A large range of feature types have been proposed for NLI and researchers have used varying combinations of these features. In §3.2 we proposed a method to measure the degree of dependence, overlap and complementarity between different features, testing it on English data. Accordingly, another aim of the present inquiry is to apply this method and quantify inter-feature diversity across multiple datasets and assess how the results compare across languages.

The final objective of thus chapter relates to estimating the upper limits of NLI accuracy. Evidence from current research indicates that some texts, particularly those of more proficient authors, can be challenging to classify and thus a pragmatic measure for such an upper-bound is needed. In §4.1 we proposed the use of an oracle for doing so, applying it to our features and the shared task data to derive realistic upper-bounds for classification accuracy. In §4.2 we applied this oracle in a cross-corpus setting and showed it can also be useful there. In this chapter we extend this oracle use to other languages and compare it with results on the English data.

The objective of these last two experiments is not solely focused on the machine learning aspects, but also relates to seeing if the same patterns we observed in previous chapters are reflected here across other languages, which is of importance to multilingual research in this area.

5.1.2 Chapter Outline

The rest of this chapter is organised as follows. We begin by presenting the corpora we use in §5.2. Our classification methodology is described in §5.3 and our feature set is outlined in §5.4. The descriptions

Target L2	Source Corpus	No. of L1 Classes	Text Length	Text Count	Topic Balanced
English	TOEFL11	11	349 (85)	12,100	Y
Spanish	ARU	6	298 (176)	206	N
Arabic	ALC	7	155 (76)	329	N
Italian	VALICO	14	210 (105)	2,531	N
German	FALKO	8	392 (135)	221	N
Chinese	JCLC	11	621 (26)	3,216	N
Finnish	LAS2	9	562 (304)	204	N

Table 5.1: A summary of the basic properties of the L2 data used in our study. The text length is the average number of tokens across the texts along with the standard deviation in parentheses.

and results from our experiments are detailed in §§5.5–5.10 and followed by a general discussion in §5.11 that summarizes the conclusions of our experiments and add some further discussion.

5.2 Data

In this section we outline the data used in this chapter. This includes the six L2 languages: in addition to outlining the corpora and their characteristics, we also describe how these languages differ linguistically and typologically from English, the most commonly investigated language in NLI.

While such corpus-based studies have become an accepted standard in SLA research and relevant NLP tasks, there remains a paucity of large-scale L2 corpora. For L2 English, the two main datasets are the International Corpus of Learner English (ICLE) (Granger, 2003) and TOEFL11 (Blanchard et al., 2013) corpora, with the latter being the largest publicly available corpus of non-native English writing.[3]

A major concern for researchers is the paucity of quality learner corpora that target languages other than English (Nesselhauf, 2004). The aforementioned data scarcity is far more acute for L2 other than English and this fact has not gone unnoticed by the research community (Lozano and Mendikoetxea, 2013; Abuhakema et al., 2008). Such corpora are few in number and this scarcity potentially stems from the costly resources required for the collection and compilation of sufficient texts for developing a large-scale learner corpus.

Additionally, there are a number of characteristics and design requirements that must be met for a corpus to be useful for NLI research, as we discussed in §2.3.1.

One key contribution of this chapter is the identification and evaluation of corpora that meet as many of these requirements as possible. The remainder of this section outlines the languages and corresponding datasets which we have identified as being potentially useful for this research and provides a summary of their key characteristics. A summary of the basic properties of the data for each language is shown in Table 5.1. Additionally, a listing of the L1 groups and text counts for each corpus can be found in Table 5.2, which also shows the average text length for documents in each class in tokens, except for Chinese, which is measured in characters.

[3] TOEFL11, previously described in §2.3.1.4, contains *c.* 4 million tokens in 12,100 texts.

5.2.1 Italian

Italian, a Romance language similar to French and Spanish, is a modern descendant of Latin. It uses the same alphabet as English, although certain letters such as *j* and *x* are only used in foreign words. As a result of being related to Latin there are various cognates between English and Italian, including a range of false friends.

Morphologically, it is a little more complicated in some ways than English. Nouns are inflected for gender (male or female) and number. However, Italian differs from other Romance languages in this regard in that the plural marker is realized as a vowel change in the gender marker and not through the addition of an *-s* morpheme. Certain nouns, such as weekdays, are not capitalized. Verbs are inflected for tense and person. In addition to the five inflected tenses, others are formed via auxiliaries. An important aspect of the verbal system is the presence of the subjunctive mood across the verb tenses. At the sentence level, SVO is the normal order, although post-verbal subjects are also allowed depending on the semantic context of the subject and verb. Pronouns are frequently dropped in Italian and this could lead to a different, possibly slightly more compact, function word distribution.

Adjectives can be positioned both pre- and post-nominally in nominal groups, with the position marking a functional aspect of its use as either descriptive (pre-nominal) or restrictive (post-nominal). This could result in a wider, more sparse distribution of POS *n*-grams. Possessives also behave in the same manner as adjectives in most contexts. A more detailed exposition of these linguistic properties, amongst others, can be found in Vincent (2009).

For our Italian data we utilise the VALICO Corpus (Corino, 2008). VALICO (*Varietà di Apprendimento della Lingua Italiana Corpus Online*, *i.e.* the Online Corpus of Learner Varieties of Italian) includes approximately 1 million tokens of learner Italian writing from a wide range of L1s along with the associated metadata.

Although over 20 native language groups are represented in the data, many do not have sufficient data for our purposes. We have selected the top 14 native languages by the number of available texts as the rest of the classes contain too few texts; these are shown in Table 5.2. In terms of the number of L1 classes, this is the highest number used in our experiments. On the other hand, there is significant imbalance in the number of texts per class.

5.2.2 German

The German language, spoken by some 100 million native speakers, is an official language of Germany, Austria, Switzerland, Luxembourg and Liechtenstein.

English and German are similar in many aspects and both belong to the Indo-European language family, as part of the West Germanic group within the Germanic branch (Hawkins, 2009).

In spite of this typological closeness, there are a number of differences that may cause problems for NLI with our standard features. German has a much richer case and morphology system compared to English and this may lead to different usage patterns of function words. Furthermore, German also has a more variable word ordering system with more long distance dependencies, potentially leading to a wider set of POS *n*-grams. It is not clear how well this feature can capture potential L1-influenced ordering patterns.

	Italian			**Chinese**	
L1	**Texts**	**Length**	**L1**	**Texts**	**Length**
Albanian	55	185	Burmese	349	618
Chinese	187	162	Filipino	415	618
Czech	84	170	Indonesian	402	619
English	310	222	Japanese*	180	621
French	335	214	Khmer	294	625
German	306	178	Korean*	330	619
Hindi	146	189	Laotian	366	630
Japanese	415	183	Mongolian	101	633
Polish	201	348	Spanish*	112	618
Portuguese	45	192	Thai	400	624
Romanian	63	207	Vietnamese	267	623
Russian	40	202			
Serbian	124	229			
Spanish	220	253			
Total	**2,531**		**Total**	**3,216**	
	Finnish			**German**	
L1	**Texts**	**Length**	**L1**	**Text**	**Length**
Czech	27	479	Chinese	11	330
English	10	653	Danish	38	442
German	21	615	English	52	475
Hungarian	21	686	French	17	512
Japanese	34	409	Polish	47	331
Komi	11	673	Russian	35	364
Lithuanian	28	634	Turkish	10	353
Polish	12	377	Uzbek	11	327
Russian	40	536			
Total	**204**		**Total**	**221**	
	Arabic			**Spanish**	
L1	**Texts**	**Length**	**L1**	**Texts**	**Length**
Chinese	76	145	English	45	326
English	35	151	French	47	410
French	44	133	German	17	325
Fulani	36	152	Greek	21	245
Malay	46	142	Italian	54	318
Urdu	64	183	Japanese	22	163
Yoruba	28	170			
Total	**329**		**Total**	**206**	

Table 5.2: A breakdown of the six languages and the L1 classes used in our study. Texts is the number of documents in each L1 class and Length represents the average text length in tokens, except for Chinese, which is measured in characters.

The largest publicly available selection of German learner texts can be found in the FALKO (***f**ehler**a**nnotierten **L**erner**ko**rpus*) corpus[4] by Siemen et al. (2006) and this is the source of the German data used in this work.

It has several sub-corpora, including the essay sub-corpus (argumentative essays written by learners) and summary sub-corpus (text summaries written by learners). It also contains baseline corpora with texts written by German native speakers. For the purposes of our experiments we combine the essay and summary texts, but do not use the longitudinal sub-corpus texts. A listing of the L1 groups and text counts of the corpus subset we use can be found in Table 5.2.

5.2.3 Spanish

As the Romance language with the greatest number of speakers, Spanish is the official language of some 20 countries.

Much like German, many aspects of Spanish grammar are similar to English, so our feature set may not have any issues in capturing L1-based interlanguage differences. Although Spanish syntax is mostly SVO it also has a somewhat richer morphology and a subjunctive mood (Green, 2009), though we do not expect these differences to pose a challenge. Pronouns are also frequently dropped and this information is captured by POS tags rather than function words. There is also a complete agreement system for number and gender within noun phrases, resulting in a wider distribution of POS n-grams. Spanish also makes pervasive use of auxiliaries, with more than fifty verbs that have auxiliary functions (Green, 2009, p. 214). This is a difference that affects distributions of both function words and POS tags.

Our Spanish learner texts were sourced from the Anglia Ruskin University (ARU) Spanish learner corpus. This is a multiple-L1 corpus[5] comprised of Spanish texts that were produced by students either as coursework or as part of exams. The texts are entered exactly as written by students and have not been corrected. The learners include undergraduates at Anglia Ruskin learning Spanish and some ERASMUS students from France, Germany and Italy. These students have varied nationalities and backgrounds (56% do not have English as L1).

Each text includes meta-data with the following information: the task set, the conditions (exam-/coursework), the text type (narrative, description, etc.), proficiency level (beginner, intermediate or advanced), course-book (where known), student identity number, L1, and gender.

A total of 20 L1s are represented in the version of the data that we received in July 2013, but many of these have too few texts to be effectively used in our experiments. Since not all the represented L1s have sufficient amounts of data, we only make use of the top six L1 categories (English, Italian, French, Japanese, Greek and German), as shown in Table 5.2.

5.2.4 Chinese

Chinese, an independent branch of the Sino-Tibetan family, is spoken by over a billion people. Unlike the other languages used in this study, Chinese orthography does not use an alphabet, but rather a logosyllabic system where each character may be an individual word or a constituent syllable.

[4] `http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko`

[5] The project under which this corpus was being compiled was never completed and the corpus was never publicly released. We were able to receive a current copy of the files from Dr. Anne Ife (`anne.ife@anglia.ac.uk`) at the Department of English, Communication, Film & Media at Anglia Ruskin University

Chinese is also an *isolating* language: there is little grammatical inflectional morphology. In contrast, other languages use inflection and auxiliaries to encode information about who did what to whom and when. In Chinese some of this information is conveyed via word order — much like in English — and an understanding of the context. Gender, number and tense may be indicated through lexical choices, or omitted entirely. More details about these unique characteristics of Chinese can be found in Li and Thompson (2009).

Levy and Manning (2003) point out three ways in which these difference may manifest themselves:

> First, Chinese makes less use of function words and morphology than English: determinerless nouns are more widespread, plural marking is restricted and rare, and verbs appear in a unique form with few supporting function words. Second, whereas English is largely left-headed and right-branching, Chinese is more mixed: most categories are right-headed, but verbal and prepositional complements follow their heads. Significantly, this means that attachment ambiguity among a verb's complements, a major source of parsing ambiguity in English, is rare in Chinese. The third major difference is subject pro-drop — the null realization of uncontrolled pronominal subjects — which is widespread in Chinese, but rare in English. This creates ambiguities between parses of subject-less structures as IP or as VP, and between interpretations of preverbal NPs as NP adjuncts or as subjects. (Levy and Manning (2003: pp. 439–440))

Given these differences, an interesting question is whether previously used features can capture the differences in the interlanguage of Chinese learners. For example, POS-based features have relied heavily on the ordering of tag sequences which are often differentiated by morphological inflections — can these features differentiate L1s in the absence of the same amount of information? The same question can be asked of function words, how does their reduced frequency affect NLI accuracy?

Growing interest has led to recent attempts at creating L2 Chinese learner corpora. One such endeavour — an ongoing project since 2006 — is being conducted at Jinan university in China, aiming to collect data from learners of Chinese. As part of the research presented in this chapter we collaborated with the researchers at Jinan University to use this data to create a machine-readable corpus of learner Chinese, suitable for research in SLA and NLP. These efforts led to the release of the Jinan Chinese Learner Corpus (JCLC), the first large-scale corpus of L2 Chinese (Wang et al., 2015). The JCLC is freely available to the research community and accessible on the web.[6] It can be used via a web-based interface for querying the data. Alternatively, the original texts can be used in text format for more advanced tasks.

Currently, the corpus contains approximately 6 million Chinese characters written by students. The majority of this data has been collected from foreign students learning Chinese at various universities in China, with some data coming from universities outside China. This data includes both exams and assignments. The texts are manually transcribed with all errors being maintained. Error annotations are not available at this stage. Learners from 59 countries are represented and proficiency levels are sampled representatively across beginner, intermediate and advanced levels. However, texts by learners from other Asian countries are disproportionately represented, with this likely being due to their geographical proximity and links to China. A more detailed description of this data can be found in Wang et al. (2015).

[6] `http://hwy.jnu.edu.cn/jclc/`

For this work we extracted 3.75 million tokens of text from the JCLC in the form of individual sentences. Following a methodology similar to that of Brooke and Hirst (2011), we combine the sentences from the same L1 to generate texts of 600 tokens on average, creating a set of documents suitable for NLI.

More specifically, the dataset composed of artificial documents is generated as follows. For each class, all the available texts are processed and the individual sentences from these texts are placed into a single pool. Once this pool has been created, we begin the process of generating artificial documents.

For each artificial text, its required minimum length is first determined by randomly picking a value within a pre-specified range $[M, N]$. This chosen value represents the minimum number of tokens or characters that are required to create a new document. By specifying this range parameter, instead of a single fixed value, we can create an artificial dataset where there is still some reasonable amount of variance in length between texts.

Sentences from the pool are then randomly allocated to the document until its length exceeds the required minimum value. The document is then considered complete; it is added to the new dataset and we proceed to generate another. It should also be noted that the document length may exceed the upper bound of the range parameter, depending on the length of the final sentence that crosses the minimum threshold. The sampling of sentences from the pool is done without replacement.

This process continues until there are insufficient sentences to create any more documents. The sentences remaining in the pool are then discarded. This procedure is performed for every class in the original dataset and yields a new dataset of artificial documents.

Manufacturing documents in this manner has a number of positive impacts. Firstly, it ensures that all documents are similar and comparable in length. If the data are being used to classify documents from another source, instead of cross-validation, the generation parameters could be changed so that the training set is similar to the test set in terms of length. Secondly, the random sampling used here means that the texts created for each class are a mix of different authorship styles, proficiencies and topics.

Although there are over 50 L1s available in the corpus, we choose the top 11 languages, shown in Table 5.2, to use in our experiments. This is due to two considerations. First, while many L1s are represented in the corpus, most have relatively few texts. Choosing the top 11 classes allows us to have a large number of classes and also ensure that there is sufficient data per-class. Secondly, this is the same number of classes used in the NLI 2013 shared task, enabling us to draw cross-language comparisons with the shared task results.

5.2.5 Arabic

Arabic, part of the Semitic family of languages, is the official language of over 20 countries. It is comprised of many regional dialects with the Modern Standard Arabic (MSA) variety having the role of a common dialect across the Arabic-speaking population.

A wide range of differences from English, some of which are highlighted below, make this an interesting test case for current NLI methods. More specifically, a rich morphology and grammar could pose challenges for syntactic features in NLI.

Arabic orthography is very different from English with right-to-left text that uses connective letters. Moreover, this is further complicated due to the presence of word elongation, common

ligatures, zero-width diacritics and allographic variants. The morphology of Arabic is also quite rich with many morphemes that can appear as prefixes, suffixes or even circumfixes. These mark grammatical information including case, number, gender, and definiteness, amongst others. This constitutes a sophisticated morphotactic system. Nouns are inflected for gender, number, case and determination, which is marked using the *al-* prefix. Verbal morphology consists of affixes for marking mood, person and aspect. For further information we refer the reader to the thorough overview in Kaye (2009).

Other researchers have noted that this morphological complexity means that Arabic has a high vocabulary growth rate, leading to issues in tasks such as language modelling (Diab, 2009; Vergyri et al., 2004). This issue could also be problematic for our POS n-gram features. Arabic function words have previously been used for authorship attribution (Abbasi and Chen, 2005) and our experiments will evaluate their utility for NLI.

The need for L1-specific SLA research and teaching material is particularly salient for a complex language such as Arabic which has several learning stages (Mansouri, 2005), such as phrasal and inter-phrasal agreement morphology, which are hierarchical and generally acquired in a specific order (Nielsen, 1997).

No Arabic learner corpora were available for a long time, but recently, the first version of the Arabic Learner Corpus[7] (ALC) was released by Alfaifi and Atwell (2013). The corpus includes texts by Arabic learners studying in Saudi Arabia, mostly timed essays written in class. In total, 66 different L1 backgrounds are represented. While texts by native Arabic speakers studying to improve their writing are also included, we do not utilize these. Both plain text and XML versions of the learner writings are provided with the corpus. Additionally, an online version of the corpus with more advanced search and browsing functionality has recently been made available.[8]

We use the more recent second version of the ALC (Alfaifi et al., 2014) as the data for our experiments. While there are 66 different L1s in the corpus, the majority of these have fewer than 10 texts and cannot reliably be used for NLI. Instead we use a subset of the corpus consisting of the top seven native languages by number of texts and as a result of this, this Arabic dataset is the smallest corpus used in an NLI experiment to date. The languages and document counts in each class are shown in Table 5.2.

5.2.6 Finnish

The final language included in this chapter is Finnish, a member of the Baltic-Finnic language group and spoken predominantly in the Republic of Finland and Estonia.

Finnish is an agglutinative language and this poses a particular challenge. In terms of morphological complexity, it is among the world's most extreme: its number of cases, for example, places it in the highest category in the comparative World Atlas of Language Structures (Iggesen, 2013). Comrie (1989) proposed two scales for characterising morphology, the index of synthesis (based on the number of categories expressed per morpheme) and the index of fusion (based on the number of categories expressed per morpheme). While an isolating language like Vietnamese would have an index of synthesis score close to 1, the lowest possible score, Finnish scores particularly high on this metric (Pirkola, 2001). Because of this morphological richness, and because it is typically associated

[7] http://www.arabiclearnercorpus.com/
[8] http://www.alcsearch.com/

with freeness of word order, Finnish potentially poses a problem for the quite strongly lexical features currently used in NLI. For more details we refer the interested reader to Branch (2009) where a detailed discussion of these characteristics is presented.

The Finnish texts used here were sourced from the Corpus of Advanced Learner Finnish (LAS2) which consists of L2 Finnish writings (Ivaska, 2014). The texts are being collected as part of an ongoing project at the University of Turku[9] since 2007 with the goal of collection suitable data than allows for quantitative and qualitative analysis of Finnish interlanguage.

The current version of the corpus contains approximately 630k tokens of text in 640 texts collected from writers of 15 different L1 backgrounds. The included native language backgrounds are: Czech, English, Erzya, Estonian, German, Hungarian, Icelandic, Japanese, Komi, Lithuanian, Polish, Russian, Slovak, Swedish and Udmurt. The corpus texts are available in an XML format and have been annotated in terms of parts of speech, word lemmas, morphological forms and syntactic functions.

While there are 15 different L1s represented in the corpus, the majority of these have fewer than 10 texts and cannot reliably be used for NLI. Instead we use a subset of the corpus consisting of the top seven native languages by number of texts. The languages and document counts in each class are shown in Table 5.2.

5.2.7 Other Corpora

In this section we describe two other corpora that we use in this chapter, specifically for Experiment 3.

5.2.7.1 The CEDEL2 Corpus

One of the first large-scale English L1–Spanish L2 corpora, the CEDEL2 corpus (Lozano, 2009; Lozano and Mendikoetxea, 2013) was developed as a part of a research project (at the Universidad Autónoma de Madrid and Universidad de Granada in Spain) that aims to investigate how English-speaking learners acquire Spanish. It contains Spanish texts written by English L1 speakers as well as Spanish native speaker controls for comparative purposes. The non-native writings are further classified into three groups according to their proficiency level (Beginner, Intermediate and Advanced). This data differs from the above-described corpora as it does not contain multiple L1 groups.

5.2.7.2 The LOCNESS Corpus

The Louvain Corpus of Native English Essays (LOCNESS)[10] — part of the Louvain family of corpora — is comprised of essays written by native English speakers. The corpus contains c. 324k tokens of text produced by British and American students. This corpus can serve as native control data for L2 English texts, given the lack of native speaker data in the TOEFL11 corpus.

Similar to the TOEFL11 data, the writings are argumentative essays on a broad set of topics ranging from nuclear power to yoga. They were written by native English speakers in universities in the US and Britain. Many of these essays are timed and collected under exam conditions, while some others are untimed. The data was collected in the early 1990s and the available meta-data indicates that these annotations may not be entirely reliable.

[9] http://www.utu.fi/fi/yksikot/hum/yksikot/suomi-sgr/tutkimus/tutkimushankkeet/las2/Sivut/home.aspx
[10] http://www.learnercorpusassociation.org/resources/corpora/locness-corpus/

5.2.8 Data Preparation Challenges

The use of such varied corpora can pose several technical and design challenges which must be addressed. Based on our experience, we list here some of issues that we encountered during these experiments, and how they were addressed.

5.2.8.1 File formats

Corpora can exist in several file formats, the most common of which are XML, HTML, word processor documents (Microsoft Word or RTF) and plain text. As a first step, it was necessary to convert all the files to a common machine-readable format before the data could be processed. We chose to convert all of the documents into a standard text format for maximum compatibility.

5.2.8.2 File Encoding

The choice of text encoding is particularly important when working with languages that use characters beyond the ASCII range. To maximize compatibility with languages and software tools, we encoded our text files as Unicode using the UTF-8 encoding without a Byte Order Mark (BOM). We also note that some languages may be represented by various character encoding standards,[11] so we found that developing programs and tools that work with a unified character encoding such as Unicode was the best way to maximize their compatibility so that they work with as many languages as possible.

5.2.8.3 Unicode Normalization

This is the process of converting Unicode strings so that all canonical-equivalent strings[12] have the exact same binary representation.[13] It may also be necessary to remove certain characters that are not part of the target language. Without the application of such safeguards, experimental results may be compromised by the occurrence of characters and symbols that only appear in texts from specific corpora or speakers of certain native languages. This effect has previously been noted by Tetreault et al. (2012) where idiosyncrasies such as the presence of characters which only appear in texts written by speakers of certain languages can compromise the usability of the corpus. This is because such characters can become strongly associated with a class and artificially inflate classification results, thus making it hard to assess the true performance of the features.

5.2.8.4 Segmentation and Tokenization

Corpora are made available with differing levels of linguistic processing. Some are pre-tokenized, some may be sentence or paragraph segmented while others are simply the raw files as produced by the authors. It is crucial to consistently maintain all files in the same format to avoid feature extraction errors. For example, if extracting character n-grams from a set of tokenized and untokenized texts, the extracted features will differ and this may influence the classification process. Accordingly, we made sure all the files had comparable formats and structures. We stored the documents with one sentence per line and each sentence was tokenized.

[11] *e.g.* GB18030, GBK and GB2312 for Chinese text

[12] In Unicode, some sequences of code points may represent the same character. For example, the character Ö can be represented by a single code point (U+00D6) or a sequence of the Latin capital letter O (U+004F) and a combining diaeresis (U+0308). Both will appear the same to a reader and are canonically equivalent, however, they will be processed as distinct features by an algorithm — hence the need to perform normalization.

[13] More information can be found at `http://www.unicode.org/faq/normalization.html`

5.2.8.5 Annotations and Idiosyncrasies

Some corpora may be annotated with additional information such as errors, corrections of errors or discourse/topical information. Where present, we removed all such information so that only the original text, as produced by the author, remained. Another minor issue has to do with the class labels that various corpora use to represent the different languages. For example, the FALKO Corpus uses the ISO 639 three letter language codes while other corpora simply use the language names or even numbers. It was necessary to create a unified set of language codes or identifiers and assign them to the texts accordingly.

5.3 Methodology

We also follow the supervised classification approach described in chapters 2 and 3. We devise and run experiments using several models that capture different types of linguistic information. For each model, features are extracted from the texts and a classifier is trained to predict the L1 labels using the features.

5.3.1 Classification

We use the very same classification approach we described in §3.2.1, *viz.* a linear SVM is used trained for each feature type using relative frequency values.

5.3.2 Evaluation

Consistent with our work thus far, most other NLI studies and the 2013 shared task, we report our results as classification accuracy under k-fold cross-validation, with $k = 10$. In recent years this has become an emergent *de facto* standard for reporting NLI results.

For creating our folds, we employ stratified cross-validation which aims to ensure that the proportion of classes within each partition is equal (Kohavi, 1995).

For comparison purposes, we define a *majority baseline*, calculated by using the largest class in the dataset as the classification output for all input documents. For example, in the case of the L2 Chinese data listed in Table 5.2, the largest L1 class is Filipino with 415 documents in a dataset with 3,216 documents in total. The majority baseline is thus calculated as $\frac{415}{3216} = 12.9\%$.

No other baselines are available here as this is the first NLI work on these corpora.

5.3.3 NLP Tools

In this section we briefly list and describe the tools used to process our data.

Chinese and German For processing these two languages, the Stanford CoreNLP[14] suite of NLP tools (Manning et al., 2014) and the provided models were used to tokenize, POS tag and parse the unsegmented corpus texts.

[14] `http://nlp.stanford.edu/software/corenlp.shtml`

Language	POS Tagset	Tag Count
Chinese	Penn Chinese Treebank Tagset	33
English	Penn Treebank Tagset	36
German	"Stuttgart/Tübinger Tagsets" (STTS)	55
Italian/Spanish	EAGLES Tagset	>300
Finnish	Custom Tagset	59

Table 5.3: A listing of the tagsets used for the languages in our experiments, including the size of the tagset.

Arabic The tokenization and word segmentation of Arabic is an important preprocessing step for addressing the orthographic issues discussed in §5.2.5. For this task we utilize the Stanford Word Segmenter (Monroe et al., 2014).[15] The Arabic texts were POS tagged and parsed using the Stanford Arabic Parser.[16]

Spanish and Italian All of the processing on these language was performed using FreeLing (Padró and Stanilovsky, 2012; Carreras et al., 2004), an open-source suite of language analyzers with a focus on multilingual NLP.

5.4 Features

From most NLI work, the majority of which we reviewed in §2.3.3, we observe that POS n-grams and function words constitute a core set of standard features for this task: these will be our fundamental features as well, as described below. We did not use any lexical features (*e.g.* word n-grams) in these experiments since none of our corpora are topic balanced and could be affected by topic bias or toponymic words, as we explained in §2.3.1.1 and §3.1.5.1.

Part-of-Speech Tags In this chapter we use POS n-grams of order 1–3, as described in §2.3.2.6. The different languages and NLP tools used to process them each utilize distinct POS tagsets. For example, our Chinese data is tagged using the Penn Chinese Treebank tagset (Xia, 2000). For Italian and Spanish, the EAGLES Tagset[17] is used while for German the Stuttgart/Tübinger Tagset (STTS) is used (Schiller et al., 1995).

A summary of these tagsets can be found in Table 5.3. Looking at these values it becomes evident that some languages have a much more detailed tag set than for other languages. It has been standard in monolingual research to just use the best available tagger and tagset that were explicitly developed for some particular language. However, this approach can be problematic in multilingual research where the tagsets, and consequently the classification results obtained by employing them, are not comparable. One possibility is to convert the tags for each language to a more general and common tagset; this would make the results more directly comparable across the languages. This issue will be further explored in Experiment IV, which is presented in §5.8.

Function Words We also use function words as features, which we described in §2.3.2.5.

[15] http://nlp.stanford.edu/software/segmenter.shtml
[16] http://nlp.stanford.edu/projects/arabic.shtml
[17] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Language	Italian	German	Spanish	Chinese	Arabic	Finnish	English
Count	399	603	351	449	150	747	400

Table 5.4: Function Word counts for the various languages in our study.

The English word list was obtained from the Onix Text Retrieval Toolkit.[18] For Chinese we compiled a list of 449 function words using Chinese language teaching resources. The complete list can be accessed online.[19] The lists for the rest of the languages have been sourced from the multilingual Information Retrieval resources made available by Prof. Jacques Savoy and can also be accessed online.[20]

The function word counts for these lists are shown in Table 5.4 and we observe that there is some variation between the list sizes across the languages. This is generally due to lexical differences and the degree of morphological complexity as the lists contain all possible inflections of the words. For example, the Finnish list contains the words *heihin*, *heille*, *heiltä*, *heissä*, *heistä* and *heitä*, all of which are declensions of the third person plural pronoun *he*. Other languages may have fewer such inflected words, leading to different list sizes.

Phrase Structure Rules Where possible, we also used Context-free Grammar Production Rules (as described in §2.3.2.7) as features. It should also be noted that the extraction of this feature is predicated upon the availability of an accurate parser for the target language. Unfortunately, this is not the case for all of our languages.

5.5 Experiment I – Evaluating Features

Our first experiment is aimed at evaluating whether the types of NLI systems and features sets employed for L2 English writings can also work for other languages. We perform NLI on the datasets of the languages described above, running experiments within each corpus, over all of the L1 classes described in §5.2.

There have been conflicting results about the optimal feature representation to use for NLI. Some have reported that binary representations perform better (Brooke and Hirst, 2012b; Wu et al., 2013) while others argue that frequency-based representations yield better results (Jarvis et al., 2013; Lahiri and Mihalcea, 2013). In §3.1.5.4 we also concluded that frequency-based feature representations worked best for the English data. This is an issue that we revisit here by comparing both representations across all of our data. This can help inform current research by determining if there are any patterns that hold cross-linguistically.

Consequently, each experiment is run with two feature representations: binary (encoding presence or absence of a feature) and normalized frequencies, where feature values are normalized to text length using the l^2-norm. We also combine the features into a single vector to create combined classifiers to assess if a union of the features can yield higher accuracy.

[18] `http://www.lextek.com/manuals/onix/stopwords1.html`
[19] `http://comp.mq.edu.au/%7Emadras/research/data/chinese-fw.txt`
[20] `http://members.unine.ch/jacques.savoy/clef/index.html`

5.5.1 Results and Discussion

The results for all of our languages are included in Table 5.5.1. The majority baseline is calculated by using the largest class as the default classification label chosen for all texts. For each language we report results using two feature representations: binary (bin) and normalized frequencies (freq).

General Observations A key finding from this experiment is that NLI models can be successfully applied to non-English data. This is an important step for furthering NLI research as the field is still relatively young and many fundamental questions have yet to be answered.

We also assess the overlap of the information captured by our models by combining them all into one vector to create a single classifier. From Table 5.5.1 we see that for each feature representation, the combined feature results are higher than the single best feature. This demonstrates that for at least some of the features, the information they capture is orthogonal and complementary, and combining them can improve results.

We also note the difference in the efficacy of the feature representations and see a clear preference for frequency-based feature values — they outperform the binary representations in all cases. Others have found that binary features are the most effective for English NLI (Brooke and Hirst, 2012b), but our results indicate the frequency representation is more useful in this task.

Below we note language-specific details and analyses.

Chinese The results show that POS tags are very useful features here. The trigram frequencies give the best accuracy of 55.60%, suggesting that there exist group-specific patterns of Chinese word order and category choice which provide a highly discriminative cue about the L1. This is interesting given that fixed word order is important in Chinese, as discussed in §5.2.4. Function word frequency features provide an accuracy of 51.91%, significantly higher than the baseline. As for English L2 texts, this suggests the presence of L1-specific grammatical and lexical choice patterns that can help distinguish the L1, potentially due to cross-linguistic transfer. We also use phrase structure rules as classification feature, achieving an accuracy of 49.80%. Again, as for English L2 data, the syntactic substructures would seem to contain characteristic and idiosyncratic constructions specific to L1 groups and that these syntactic cues strongly signal the writer's L1.

The Chinese data is the largest corpus used in this work and also has the same number of classes as the TOEFL11 corpus used in the 2013 NLI shared task. This enables us to compare the results across the datasets to see how these features perform across languages. However there are also a number of caveats to bear in mind: the corpora differ in size and the Chinese data is not balanced by class as TOEFL11 is. We perform the same experiments on TOEFL11 using the English CoreNLP models, Penn Treebank POS tagset and our set of 400 English function words. Figure 5.1 shows the results side by side.

Perhaps surprisingly, we see that the results closely mirror each other across corpora in terms of relative strengths of feature types. This may be connected to the strongly configurational nature of both English and Chinese.

Arabic The frequency distributions of the production rules yield 31.7% accuracy and function words achieve 29.2%. While all the models provide results above the baseline, POS tag n-grams are

Feature	Chinese		Arabic		Italian		Finnish		German		Spanish	
	Bin	Freq	Bin	Freq	Bin	Freq	Bin	Freq	Bin	Freq	Bin	Freq
Maj. Baseline	12·90	12·90	23·10	23·10	16·40	16·40	19·61	19·61	23·53	23·53	26·21	26·21
Func. Words	43·93	51·91	22·70	29·20	45·27	50·10	46·97	54·60	49·32	54·59	42·71	47·44
POS unigrams	20·12	35·32	24·50	36·04	42·59	48·28	23·57	36·30	32·27	38·09	38·52	42·23
POS bigrams	32·83	54·24	28·33	37·60	49·17	54·03	35·82	55·20	44·49	48·41	43·77	45·16
POS trigrams	47·24	55·60	29·14	36·50	54·26	56·89	37·09	54·80	45·40	50·22	48·15	51·42
Prod. Rules	36·14	49·80	24·18	31·70	–	–	–	–	–	–	–	–
All Combined	61·75	70·61	37·08	41·00	65·82	69·09	52·49	58·86	57·12	60·32	51·32	56·18

Table 5.5: NLI classification accuracy (%) for Chinese (11 classes), Arabic (7 classes), Italian (14 classes), Finnish (9 classes), German (8 classes) and Spanish (6 classes). Results are reported using both binary and frequency-based feature representations. The production Rules features were only tested on some languages.

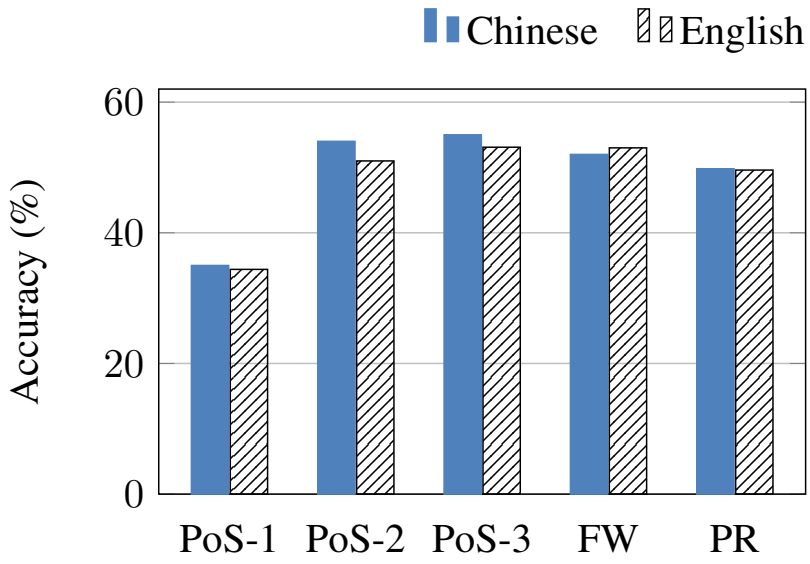

Figure 5.1: Comparing feature performance on the CLC and TOEFL11 corpora. POS-1/2/3: POS uni/bi/trigrams, FW: Function Words, PR: Production Rules.

the most useful features. Combining all of the models into a single feature space provides the highest accuracy of 41%.

The Arabic results deviate from the other language in several ways. First, the improvement over the baseline is much lower than for other languages. Second, although POS bigrams provide the highest accuracy for a single feature type with 37.6%, this is very similar to the POS unigrams and trigrams. Production rules were also worse than POS n-grams. Third, although the combined model is higher than the single-feature models, this is a much smaller boost compared to other languages. All of these issues could potentially be due to data size.

The Arabic data is our smallest corpus, which to the best of our knowledge, is the smallest dataset used for NLI in terms of document count and length. In this regard, we are surprised by relatively high classification accuracy of our system, given the restricted amount of training data available. While it is hard to make comparisons with most other experiments due to differing number of classes, one comparable study is that of Wong and Dras (2009b) which used some similar features on a 7-class English dataset. Despite their use of a much larger dataset,[21] our individual models are only around 10% lower in accuracy.

In their study of NLI corpora, Brooke and Hirst (2011) showed that increasing the amount of training data makes a very significant difference in NLI accuracy for both syntactic and lexical features. This was verified by Tetreault et al. (2012) who showed that there is a very steep rise in accuracy as the corpus size is increased towards 11,000 texts.[22] Based on this, we expect that given similarly sized training data, an Arabic NLI system can achieve similar accuracies.

Italian We make use of all 14 classes available in the VALICO corpus. This is the largest number of classes in this work and one of the highest to be used in an NLI experiment.[23] Function words scored 50.1% accuracy and POS trigrams yielded 56.89%. Combining all of the features together improves this to 69.09%, four times higher than the baseline.

[21] Wong and Dras (2009b) had 110 texts per class, with average text lengths of more than 600 words.
[22] Equivalent to 1000 texts per L1 class.
[23] Previously, Torney et al. (2012) used 16 classes from the ICLE corpus.

Finnish Here we observe that the distribution of function words yields 54.6% accuracy. This is perhaps unexpected in that Finnish, as a morphologically rich language, has a reduced role for function words relative to other languages. We believe their usefulness here is due to the use of an IR stoplist which contains more than just linguistically defined closed-class words.

The best single-feature accuracy of 54.8% comes from POS trigrams. This may also be unexpected, given that Finnish has much freer word order than the other languages in this study. But the gap over function words is only 0.2%, compared to the strongly configurational Chinese and Italian, where the gap is 4–6%.

The combined model provides the highest accuracy of 58.86%, around 4% better than the best single feature type. An interesting difference is that POS unigrams achieve a much lower accuracy of 36.3%.

German Here function words are the best single feature for this language. This deviates from the results for the other languages where POS n-grams are usually the best syntactic feature. Again, this may reflect the nature of the language: German, like Finnish, is not (strongly) configurational.

Spanish The pattern here is most similar to Chinese and Italian.

5.5.2 Learning Curves

We can also examine the learning curves of these features across various languages to see if the learning rates differ cross-linguistically.

These curves are generated by incrementally increasing the size of the training set, from 10% through to 90%. We produce one curve for each feature-language pair, using two of the best performing features: Function Words and POS trigrams. As a dataset needs to be sufficiently large for training on 10% of it to give a meaningful result, we analyze the curves only for the English TOEFL11 data and our two biggest non-English datasets: Chinese and Italian. The curves are presented in Figure 5.2.

The curves demonstrate similar rates of learning across languages. We also note that while the relationship between function word and POS trigram features isn't perfectly constant across number of training examples, there are still discernible trends. The English and Chinese data are most suitable for direct comparison as they have the same number of classes. Here, we see that Function Words provide similar accuracy scores across both languages with 2000 training documents. They also plateau at a similar score. Similar patterns can be observed for POS trigrams.

5.6 Experiment II – Comparing Languages

The focus of our second experiment is to systematically compare the performance of our feature set across a range of languages. Here, we are interested in a more direct cross-linguistic comparison on datasets with equal numbers of classes. We approach this by using subsets of our corpora so that they all have the same number of number of classes.

We run this experiment using our two biggest corpora, Chinese and Italian. Additionally, we also compare our results to a subset of the TOEFL11 L2 English corpus. Table 5.6 shows the six languages that were selected from each of the three corpora. The number of documents within each class are

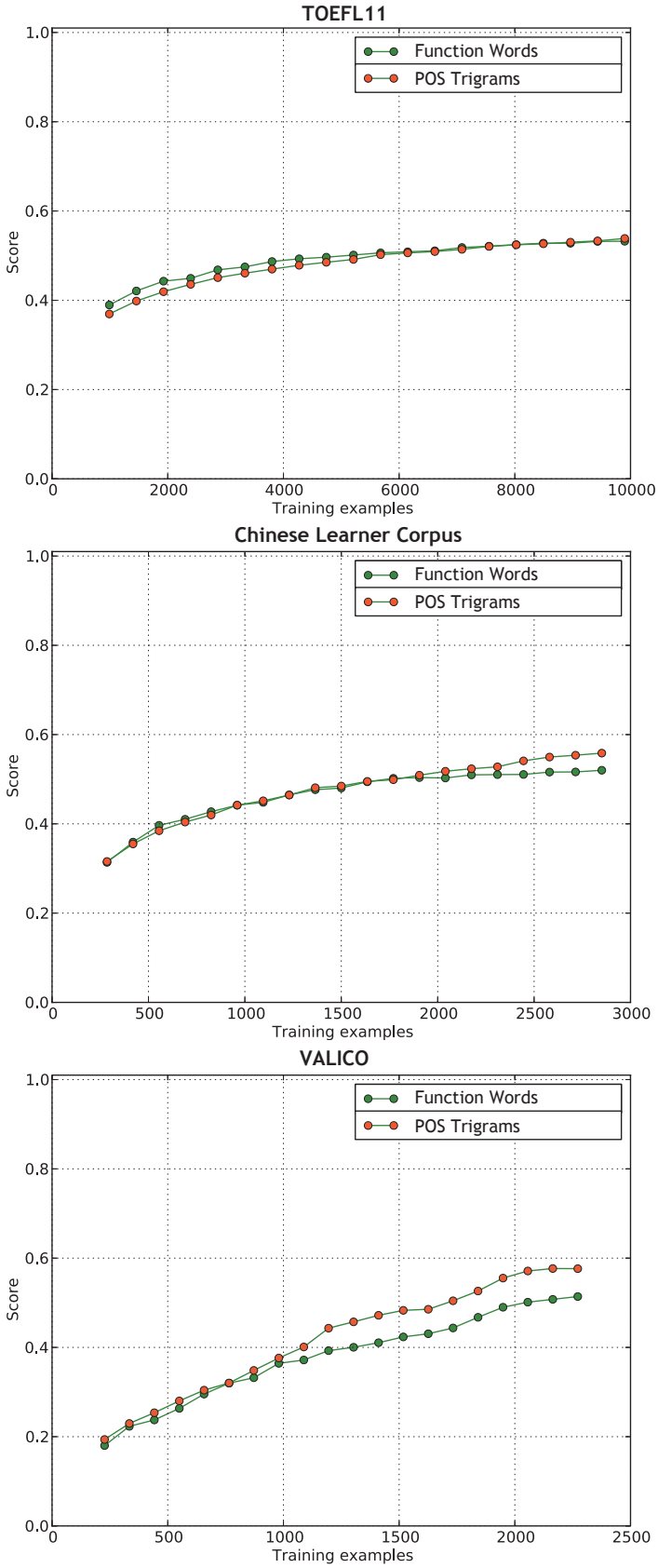

Figure 5.2: The learning curves (classification accuracy score vs. training set size) for two feature types, Function Words and POS trigrams, across three languages: English (TOEFL11, row 1), Chinese (CLC, row 2) and Italian (VALICO, row 3).

Language	L1 Classes
Chinese	Filipino, Indonesian, Thai
330 texts per class	Laotian, Burmese, Korean
Italian	French, Japanese, Spanish
200 texts per class	English, Polish, German
English	French, Japanese, Spanish
1,100 texts per class	Hindi, Turkish, Arabic

Table 5.6: The six L1 classes used for each language in Experiment II.

Feature	Accuracy		
	Chinese	Italian	English
Random Baseline	16·67%	16·67%	16·67%
(1) Function Words	62·12%	59·24%	63·82%
(2) POS unigrams	47·78%	51·15%	48·59%
(3) POS bigrams	63·14%	63·58%	63·70%
(4) POS trigrams	64·31%	64·66%	65·62%
All features (1–4)	68·14%	67·61%	70·05%

Table 5.7: Comparing classification results across languages.

kept even. These languages were chosen in order to maximize the number of classes and the number of documents within each class.

Given that the results of Experiment I favored the use of frequency-based feature values, we also use them here. We anticipate that the results will be higher than the previous experiment, given that there are fewer classes.

5.6.1 Results

The results for all three languages are shown in Table 5.7. Each language has a majority class baseline of 16.67% as the class sizes are balanced. The results follow a similar pattern as the previous experiments with POS trigrams being the best single feature and a combination of everything achieving the best results.

Chinese yields 68.14% accuracy, while Italian and English data obtain 67.61% and 70.05%, respectively. All of these are more than 4 times higher than the baseline.

5.6.2 Discussion

These results, shown graphically in Figure 5.3, demonstrate very similar performances across three different L2 corpora, much like the results in Experiment 1 for comparing English and Chinese performances. The results are particularly interesting as the features are performing almost identically across entirely different L1–L2 pairs. Again, as in §5.5.1, this may be related to the degree of configurationality in these languages.

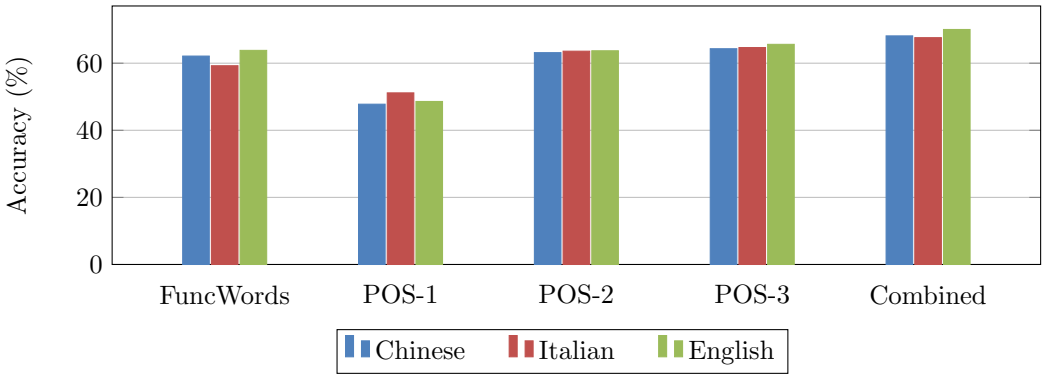

Figure 5.3: Performance of our syntactic features (Function Words and Part-of-Speech 1-3 grams) across the three languages.

Here we also see that combining the features provides the best results in all cases. We also note that the English data is much larger than the others. It contains a total of 6,600 texts (evenly distributed with 1,100 per language) and this is a probably reason for the slightly higher performance.

5.7 Experiment III – Identifying Non-Native Writing

Our third experiment involves using the previously described features to classify texts as either having been written by a native speaker (NS) or non-native speaker (NNS) author. This should be a feasible task, given the results of the previous experiments. The objective here to see to what degree our features can distinguish the writings of non-native speakers and how this performance varies across the three different languages for which we have NS data: Finnish, Spanish and English.

We approach this in a similar manner to the previous experiments, with the exception that this is a binary classification task for distinguishing two classes: native speaker author (NS) and non-native speaker author (NNS). Texts for the NNS class will come from learner corpora of three different languages while data from native speaker controls is used for the NS class, as we describe here.

For Finnish we utilize a set of 100 control texts included in the LAS2 corpus that are written by native Finnish speakers. These represent the NS class. This is contrasted against the NNS class which includes 100 texts in total, sampled as evenly as possible[24] from each language[25] listed in Table 5.2.

For Spanish we use the CEDEL2 corpus, described in §5.2.7.1. Here we use 700 native speaker texts along with another set of 700 NNS texts randomly drawn from the essays of L1 English speakers. All texts are sourced from the same corpus and have a similar topic distribution.

Finally, we also apply these methods to L2 English data using the Toefl11 and LOCNESS corpora, described in §2.3.1.4 and §5.2.7. This is required as the Toefl11 corpus does not contain any native control texts. The NS class is composed of 400 native speaker essays taken from the LOCNESS corpus and the NNS data comes from the Toefl11 corpus. We sample this data evenly from the 11 L1 non-native classes, selecting 36 or 37 texts from each to create a total of 400 texts.

The number of documents in both classes for each language are equal, hence all results are compared against a random baseline of 50%. This experiment only uses frequency-based feature value representations and results are reported as classification accuracy under 10-fold cross-validation.

[24] So that the non-native speaker class consists of a similar number of texts from each L1 class.

[25] English only has 10 texts, so we include 2 extra Japanese texts to create a set of 100 documents, with roughly 11 texts from each L1 class.

Feature	Accuracy		
	Finnish	Spanish	English
Random Baseline	50·00%	50·00%	50·00%
(1) Function Words	93·96%	91·12%	94·26%
(2) Part-of-Speech unigrams	88·54%	88·71%	87·91%
(3) Part-of-Speech bigrams	90·15%	90·35%	91·81%
(4) Part-of-Speech trigrams	91·45%	91·35%	92·87%
(5) Production Rules	N/A	91·28%	93·61%
All features combined	**94·92%**	**95·23%**	**96·45%**

Table 5.8: Accuracy for classifying texts as Native or Non-Native.

5.7.1 Results and Discussion

Table 5.7.1 shows the results for all three languages, demonstrating that all features greatly surpass the 50% baseline for all languages. The use of function words is the best single feature for two of the three languages, but combining all the features provides the best accuracy of approximately 95% in all cases.

Our Finnish data is relatively small with 100 documents in each class, thus we see that our commonly used features are largely sufficient for this task, even on a small dataset. The combined model achieves an accuracy of 94.92%.

For Spanish, all features with the exception of POS unigrams achieve accuracies of over 90%. When combined, the model yields the best accuracy of 95.23%.

The Spanish and Finnish results are very similar, despite Spanish having a much larger dataset. To investigate this further, we examined the learning curve for the best Spanish feature — POS trigrams — as shown in Figure 5.4.

We note that although the accuracy increases as the amount of data increases, the curve is much flatter than those for NLI in §5.5. However, this is offset by the fact that the curve's starting point is much higher, achieving over 85% accuracy by using only 10% of the data for training.

The English results are very similar to those from the other languages and the combined model scores a cross-validation accuracy of 96.45%. The texts in the English experiment here — unlike the Finnish and Spanish data — are sourced from different corpora, but while they are all student essays, they may differ significantly in topic, genre and style. some of these differences were highlighted in §5.2.7.2. This was a limitation that we were unable to overcome with the currently available data. In future work, further topic-controlled experiments can also be performed for English using a dataset that contains sufficient amounts of native and non-native data for the same topic. The use of synthetic data could have an impact on such studies. Ideally, we would have a dataset that was collected under the same exam conditions as the TOEFL11 and on the same topics. However, to the best of our knowledge, no corpus of timed, exam-condition essays written by native English speakers is currently available to the research community.

One direction for future experiments is the investigation of the relationship between L2 proficiency and the detection accuracy of non-native writing. Previous results by Tetreault et al. (2012) show that NLI accuracy decreases as writing proficiency improves and becomes more native-like, but would

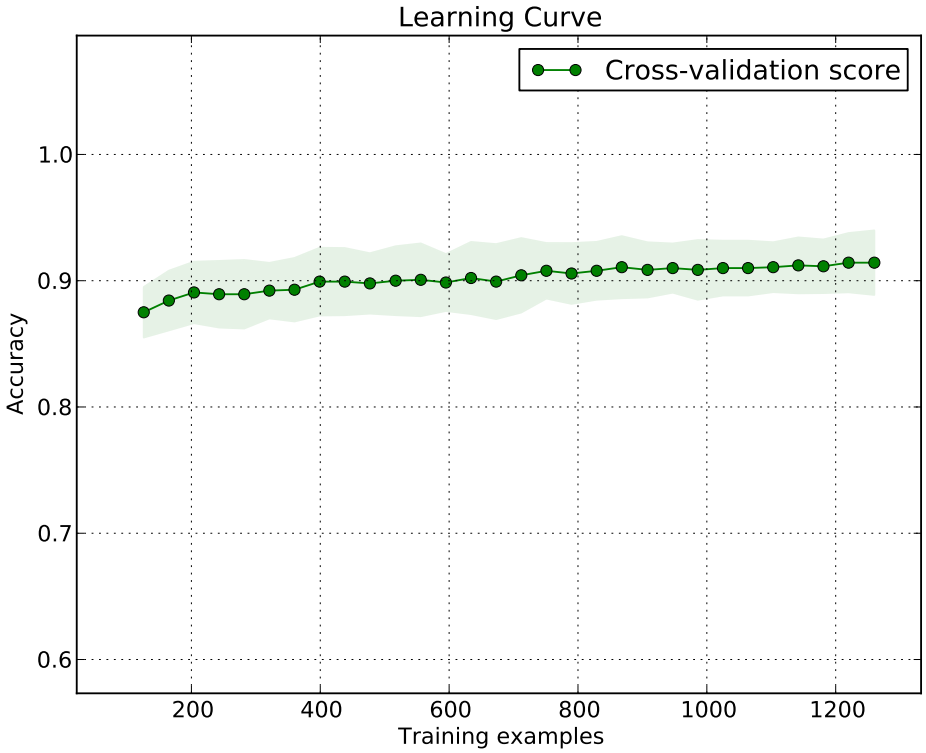

Figure 5.4: A learning curve for the Spanish Native vs. non-Native classifier trained on POS trigrams. The standard deviation range is also highlighted.

this pattern hold here as well? Another potential path for future work is to extend this experiment to the sub-document level to evaluate the applicability of this approach at the paragraph, or even sentence level.

It should also be noted that it may be possible to approach this problem as a verification task (Koppel and Schler, 2004) instead of a binary classification one. In this scenario the methodology is one of novelty or outlier detection where the goal is to decide if a new observation belongs to the training distribution or not. This can achieved using one-class classifiers such as a one-class SVM (Schölkopf et al., 2001). One option is select native writing as the inlier training class as it is easier to characterize native writing and more importantly, training data is more readily available. It is also harder to define non-native writing as there can be many varieties, as we have shown in our experiments thus far. This is something we aim to investigate in future work by comparing the two approaches.

5.8 Experiment IV – The effects of POS tagset size on NLI accuracy

POS tagging is a core component of many NLP systems and this is no different in the case of NLI, as evidenced by experimental results thus far. Over the last few decades, a variety of tagsets have been developed for various languages and treebanks. Each of these tagsets is often unique and tailored to the features of a specific language. Within the same language, the existing tagsets can differ in their level of granularity.

Tagsets differ in size according to their level of syntactic categorization which provides different levels of syntactically meaningful information. They can be very fine-grained in their distinction between syntactic categories by including more morpho-syntactic information such as gender, number, person, case, tense, verb transitivity and so on. Alternatively, a more coarse-grained tagset may only use broader syntactic categories such as verb or noun. This can be observed by looking at some of the tagsets developed for English, *e.g.*:

Penn Treebank Tagset (Marcus et al., 1993) – 36 tags

Brown Corpus Tagset (Greene and Rubin, 1971) – 87 tags

CLAWS2 Tagset (Garside, 1987) – 166 tags

SUSANNE Corpus Tagset (Sampson, 1993) – 352 tags

In this chapter we also made use of a slew of different tagsets for the different languages, which were outlined in §5.4. Since the n-grams extracted from these POS tags can help capture characteristic word ordering and error patterns, it could be argued that a larger tagset can generate more discriminative sequences and thus yield better classification performance. However, it can also result in much larger and more sparse feature vectors.[26] Using the Penn Treebank and CLAWS2 tagsets, in §3.1 we concluded that a more linguistically fine-grained POS tagset produced better results for English NLI.

Accordingly, the aim of this experiment is to assess the effect of POS tagset size on classification accuracy, hypothesizing that a larger target will provide better results. We base this on our initial results from §3.1.3.5 which showed that a larger tagset achieved better accuracy. We also aim to compare the effectiveness of POS tags cross-linguistically. This could also enable us to better understand the results from §5.5 and what impact the different granularity of tagsets might have had.

Some previous research has examined this issue on L2 English data, but no complete comparison is available. Gyawali et al. (2013) report that the use of a smaller tagset reduced English NLI accuracy. To further investigate this issue, we conduct a more thorough, cross-linguistic comparative evaluation of tagset performance.

5.8.1 A Universal Part of Speech Tagset

While a number of different tagsets have been proposed, certain tasks such as cross-lingual POS tagging (Täckström et al., 2013), multilingual parsing (McDonald et al., 2013) or drawing comparisons across tagsets require the use of a common tagset across languages. To facilitate such cross-lingual research, Petrov et al. (2012) propose a Universal POS Tagset consisting of twelve coarse POS categories that are considered to be universal across languages.[27]

We utilize this Universal POS Tagset (UPOS) in this experiment and convert the tags in the three largest datasets available: English, Chinese and Italian. By mapping from each language-specific tagset to the universal one, we obtain POS data in a common format across all languages. This enables us to compare the relative performance of the original and reduced tagset data. It also

[26] Theoretically the n-gram space grows exponentially with n, but the growth rate is lower in practice as not all possible combinations are observed in the data.

[27] These categories are: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), "." (punctuation marks) and X (a catch-all for other categories such as abbreviations or foreign words). These categories were derived through analysis of tagsets proposed for 22 different languages.

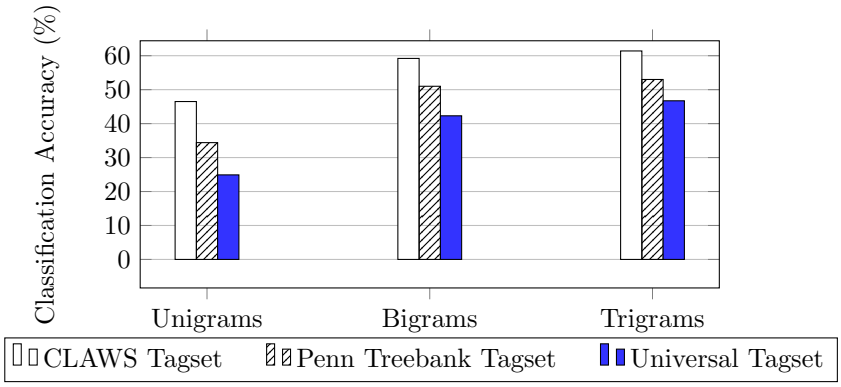

Figure 5.5: NLI classification accuracy for L2 English data from the TOEFL11 corpus, using POS n-grams extracted with the CLAWS, Penn Treebank and Universal POS tagsets.

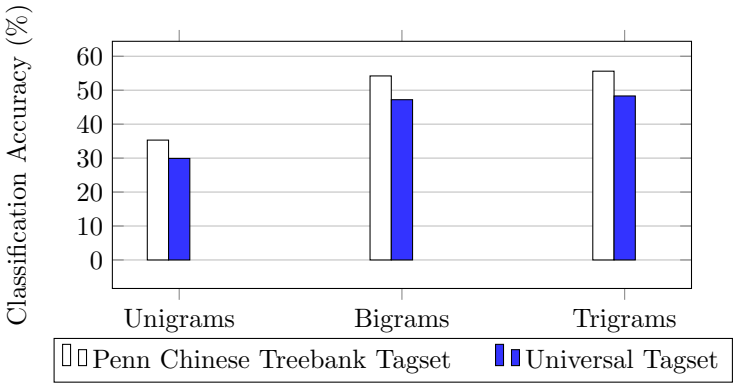

Figure 5.6: NLI classification accuracy for the L2 Chinese data, using POS n-grams extracted with the Penn Chinese Treebank and Universal POS tagsets.

permits us to compare the utility of POS tags as a classification feature across languages. For English we experiment with three tagsets: CLAWS, Penn Treebank (PTB) and UPOS.

5.8.2 Results and Discussion

The results for English are shown in Figure 5.5 and demonstrate that the largest tagset — CLAWS — provides the best classification accuracy. Classification accuracy continues to drop as the tagset gets smaller.

Figures 5.6 and 5.7 show the results for Chinese and Italian, respectively. Here we see a similar pattern, but the performance drop is much steeper for Italian. This is likely because the Italian data uses a much more fine-grained tagset than Chinese.[28]

A notable finding here, related to our first hypothesis, is that larger tagsets always yield higher classification results. Evidence from all three languages supported this.

However, these results also show that even with only 12 POS tags, the UPOS set retains around 80% of the classification accuracy of the full tagsets. This finding signals that the great majority of the syntactic patterns that are characteristic of L1 groups are related to the ordering of the most basic word categories. This can be further investigated by comparing learner data with the same L1 but multiple L2s[29] to find common transfer patterns related to that L1.

[28] We observe 330 tags in our Italian data while the Penn Chinese Treebank only uses 33 tags. The reduction from 330 to 12 tags is steeper, hence the greater drop in accuracy.

[29] *e.g.* comparing Chinese-English, Chinese-Spanish and Chinese-French.

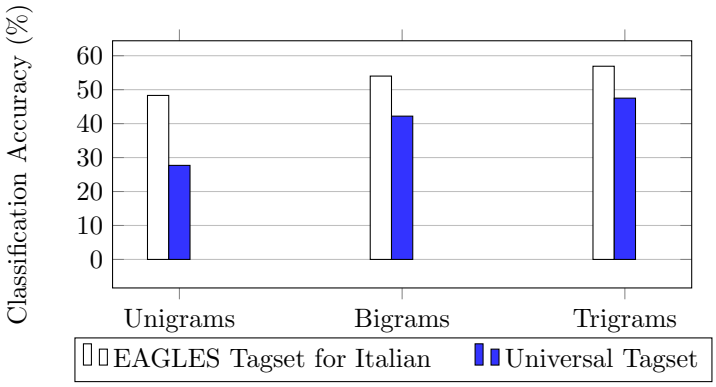

Figure 5.7: NLI classification accuracy for the L2 Italian data using POS n-grams extracted with the EAGLES and the Universal POS tagsets.

Another interesting observation is that the UPOS results are quite similar and closely mirror each other across the three languages. *Prima facie*, this supports previous findings suggesting that a systematic pattern of cross-linguistic transfer may exist, where the degree of transfer is independent of the L1 and L2 (Malmasi and Dras, 2014b). While these results are certainly not conclusive, this is a question that merits further investigation, pending the availability of additional learner corpora in the future.

Finally, as evidenced by our results, we can also conclude that the use of a universal tagset can be helpful in comparing the performance of syntactic features such as POS tags in cross-lingual studies where the languages use distinct tagsets.

5.9 Experiment V – Bounding Classification Accuracy

In Chapter 4 we proposed the use of oracles for measuring the upper bound an classification accuracy for a dataset. The aim of this experiment to evaluate this methodology on additional data beyond English.

In this experiment we use the selected feature set on our biggest datasets: Chinese, Italian and English. Following the oracle methodology described in §4.1.1, we train a single classifier for each feature type to create our ensemble. We do not experiment with combining different machine learners here; instead we focus on gauging the potential of the feature set.

The oracle classifier fusion method is then run on each ensemble so that the correct label is assigned to each document if any of the classifiers in the ensemble classify it correctly. These labels are then used to calculate the potential accuracy of the ensemble on the dataset. We perform this procedure for each language.

5.9.1 Results

The oracle results for the three languages are shown in Table 5.9 and contrasted against the majority class baseline and our combined features classifier. These results establish that NLI systems have the potential to achieve high classification accuracy. Analyzing the relative increase over the baseline shows better performance on larger datasets.

The results indicate that at least one of our feature types was able to correctly classify some 85% of the texts in each dataset. However, even under this best scenario, we should note that not a single

	Italian	Chinese	English
Majority Baseline	16·40%	12·90%	09·09%
Our Best Accuracy	69·09%	70·61%	70·58%
Oracle Accuracy	84·35%	87·60%	86·20%

Table 5.9: Oracle classifier accuracy for the three languages in experiment V.

classifier is able to correctly predict the label for the remaining 15% of the data. This suggests that a certain portion of L2 texts are not distinguishable by any of our current features. This value is similar across the three languages, indicating that this may be a more general trend.

5.9.2 Discussion

This experiment evaluated an oracle classifier for estimating the "potential" upper limit of NLI accuracy on several datasets. This upper limit can vary depending on which components — feature types and algorithms — are used to build the NLI system. Alongside other baseline measures, the Oracle performance can be helpful in interpreting the relative performance of an NLI system.[30]

A useful application of this method is to isolate the subset of wholly misclassified texts for further investigation and error analysis. This segregated data can then be independently studied to better understand the aspects that make it hard to classify them correctly. This can also be used to guide feature engineering practices in order to develop features that can distinguish these challenging data points.

As the oracle accuracy is similar across the three languages, this may indicate a more general trend related to writing proficiency and the maximum classification potential. Previously, the work of Tetreault et al. (2012) demonstrated that classification gets increasingly harder as writer proficiency increases. This higher proficiency makes it more challenging to discern the native-like writings of authors of distinct L1 backgrounds. It may also point to a deficiency in the feature set: a portion of the data are indistinguishable using the current features.

5.10 Analyzing Feature Diversity

Earlier in §3.2 we outlined a method for analyzing feature diversity and applied the method to English data. In this section we complement those initial results by running the same analysis for Chinese and Arabic and comparing them with English to see if similar diversity patterns hold across different languages.

5.10.1 Results

We calculate the Q coefficient for our largest dataset, Chinese, using all five features listed in Table 5.5.1. For comparison purposes, we also calculate Q for the same features on the Arabic and TOEFL11 English data. The matrices of the Q coefficients for all features and languages are shown graphically in Figure 5.8. We did not find a negative correlation between any of our features.

[30] *e.g.* an NLI system with 70% accuracy against an Oracle baseline of 80% is relatively better compared to one with 74% accuracy against an Oracle baseline of 93%.

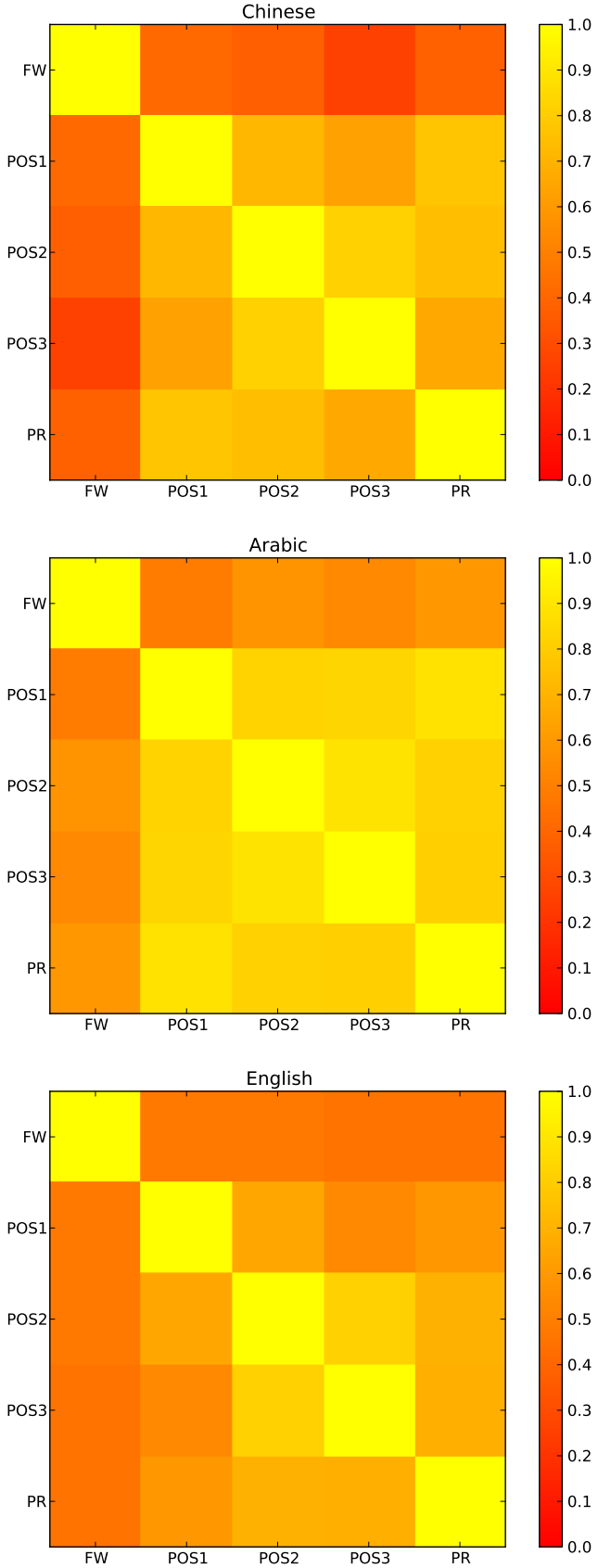

Figure 5.8: The Q-coefficient matrices of five features for Chinese (top), Arabic (middle) and English (bottom). The matrices are displayed as heat maps. POS 1/2/3: POS uni/bi/trigrams, FW: Function Words, PR: Production Rules.

The values for Chinese show a weak correlation of 0.3 between Function Words and all other features. Production Rules also have a moderate correlation with POS trigrams. Additionally, although their outputs are weakly to moderately correlated, these three features yield similar accuracy when used independently. Such features, with high individual accuracy yet low output correlation are ideal sources of diversity when combining classifiers.

Looking at the other languages, we also observe very similar patterns across the data, as can be seen by comparing the plots in Figure 5.8. This seems to suggest that these correlation patterns may hold cross-lingually.

To test the validity of these results, we re-ran the Chinese experiment from §5.5, this time combining the top 3 features with the lowest average Q coefficient, weighted by their classification error.[31] These features are Function Words, Production Rules and POS trigrams and combining them yields an accuracy of 70.7%, compared to 70.6% for using all five features. This, then, suggests that the most diverse features contribute the most to the combined classifier and that removing redundant information can increase accuracy. Having several highly dependent feature types may make it harder for a learner to overcome their errors.

5.11 General Discussion

The study presented in this chapter examined a number of different issues from a cross-lingual perspective, making a number of novel contributions to NLI research. Using up to six languages to inform our research, our experiments use evidence from multiple languages to support their results and to identify general patterns that hold across multiple languages. The most prominent finding here is that NLI techniques can be successfully applied to a range of languages that differ from English, which has been the focus of almost all previous research.

To the best of our knowledge this is the first sizeable study of NLI with a primary focus on multiple non-English L2 corpora. This includes the identification of relevant data and tools for conducting cross-lingual NLI research. We believe this is an important step for furthering NLI research as the field is still relatively young and many fundamental questions remain unanswered. These results are useful for gaining deeper insights about the technique and exploring its potential application in a range of contexts, including education, SLA and forensic linguistics.

Our first two experiments evaluated our features and data, showing that the selected commonly used features perform well and with similar accuracy across languages, taking into account corpus size; they also suggest that the effectiveness of particular feature types is related to the typological character of the L2 to some extent. The third experiment compared non-native and native control data, showing that they can be discerned with 95% accuracy. We also looked at multilingual evaluations of other NLI issues. Experiment V applied an oracle to several languages, aiming to calculate the upper limits of classification accuracy in this task. Our feature diversity analysis in §5.10 also showed that feature dependence patterns hold across several languages.

We also note the difference in the efficacy of the feature representations and see a clear preference for frequency-based feature values. Others have found that binary features are the most effective for English NLI (Brooke and Hirst, 2012b), but our results indicate frequency information is more informative in this task.

[31] For feature i this is calculated as $\bar{Q}_i \times (1 - \text{Accuracy}_i)$; lower values suggest higher accuracy and diversity.

Additionally, the corpora we have identified here can be used in other NLP tasks, including error detection and correction. This, of course, depends largely on the kinds of annotations the corpora have. If not already present, the corpora would need to be annotated for grammatical errors and their corrections.

There are also a number of methodological shortcomings that merit discussion. In its current state, research in non-English NLI is affected by many of the same issues that were prevalent in English NLI research prior to the release of the TOEFL11 corpus. This includes the lack of a common evaluation framework and a paucity of large-scale datasets that are controlled for topic, the number of texts across the various L1 classes and also text length.

This study is affected by many such issues, *e.g.* a lack of even amounts of training data, as none of the non-English corpora used here were designed specifically for NLI. However, it should be noted that many of the early studies in English NLI were performed under similar circumstances. These issues were noted at the time, but did not deter researchers as corpora with similar issues were used for many years. Non-English NLI is also at a similar state where the extant corpora are not optimal for the task, but no other alternatives exist for conducting this research.

In addition to this data paucity, the lack of NLP tools for all languages is another limiting factor that hinders further research. Many aspects of NLI studies require the use of accurate parsers and taggers to extract relevant information from learner texts. The set of features used in this work was limited by the availability of linguistic tools for our chosen languages.

Finally, we would also like to point to the failure to distinguish between the L2 and any other acquired languages as a more general criticism of the NLI literature to date. The current body of NLI literature fails to distinguish whether the learner language is in fact the writer's second language, or whether it is possibly a third language (L3). None of the corpora used here contain this metadata.

It has been noted in the SLA literature that when acquiring an L3, there may be instances of both L1- and L2-based transfer effects on L3 production (Ringbom, 2001). Studies of such second language transfer effects during L3 acquisition have been a recent focus in cross-linguistic influence research (Murphy, 2005).

One potential reason for this shortcoming in NLI is that none of the commonly used corpora distinguish between the L2 and L3; they only include the author's L1 and the language being learned. This language is generally assumed to be an L2, but may not be case. At its core, this issue relates to corpus linguistics and the methodology used to create learner corpora. The thorough study of these effects is contingent upon the availability of more detailed language profiles of authors in learner corpora. The manifestation of these interlanguage transfer effects (the influence of one non-native language on another) is dependent on the status, recency and proficiency of the learner's acquired languages (Cenoz and Jessner, 2001). Accordingly, these variables need to be accounted for by the corpus creation methodology.

It should also be noted that based on currently available evidence, identifying the specific source of cross-linguistic influence in speakers of an L3 or additional languages (L4, L5, etc.) is not an easy task. Recent studies point to the methodological problems in studying productions of multilinguals (De Angelis, 2005; Williams and Hammarberg, 1998; Dewaele, 1998).

From an NLP standpoint, if the author's acquired languages or their number is known, it may be possible to attempt to trace different transfer effects to their source using advanced segmentation

techniques. We believe that this is an interesting task in itself and a potentially promising area of future research.

Although specific directions for future research were discussed within each experiment, there are also a number of broader avenues for future work. The extension of these experiments to additional languages is the most straightforward direction for future research. The goal here would be to verify if the trends and patterns found in this work can be replicated in other languages. This can be expanded to a more comprehensive framework for comparative studies using equivalent syntactic features but with distinct L1–L2 pairs to help us better understand Cross-Linguistic Influence and its manifestations. Such a framework could also help us better understand the differences between different L1–L2 language pairs.

The potential expansion of the experimental scope to include more linguistically sophisticated features also merits further investigation, but this is limited by the availability of language-specific NLP tools and resources. Such features include dependency parses, language models, stylometric measures and misspellings. The cross-lingual comparison of these features may identify additional trends.

A common theme across the first three experiments was that the combination of features provided the best results. This can be further extended by the application of classifier ensemble methods. This could be done by aggregating the output of various classifiers to classify each document, similar to the work of Tetreault et al. (2012) for English NLI. We discussed a number of such ensemble combination methods in §3.1. The methods described in our feature diversity analysis from §5.10 could also help guide the selection of diverse features to reduce redundancy in the classifier committee.

5.12 Chapter Summary

In this chapter we:

- identified sources of suitable NLI data for six typologically different languages
- released the Jinan Chinese Learner Corpus
- applied NLI to these languages, achieving similar performance for the six L2s
- established that performance of common features is similar across different languages
- concluded that more fine-grained POS tagsets also perform better for other languages
- demonstrated that oracle performance patterns are similar for other L2s
- showed that feature diversity trends are similar across languages
- evaluated the classification of native and non-native writing, achieving high accuracy

Following this, in the next chapter we:

- examine how NLI models can inform SLA research

Chapter 6

Extracting Language Transfer Hypotheses and Error Contexts

Having established that NLI as a classification task is applicable to a range of L1s, in this chapter we look at how information captured by models from this task could potentially inform SLA research. We do this in two parts. We first examine the extracted features to see if there is any potential for using the various linguistic feature types in SLA. In the second part we look at how SLA hypotheses might be constructed using such features.

Chapter Contents

The work presented in this chapter has been published as:

- Shervin Malmasi and Mark Dras. Language Transfer Hypotheses with Linear SVM Weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1385–1390, Doha, Qatar, October 2014d. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D14-1144`
- Shervin Malmasi and Mark Dras. From Visualisation to Hypothesis Construction for Second Language Acquisition. In *Proceedings of TextGraphs-9: the Workshop on Graph-based Methods for Natural Language Processing*, pages 56–64, Doha, Qatar, October 2014f. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W14-3708`

6.1 Hypotheses in SLA

In §2.4 we briefly discussed the field of Second Language Acquisition (SLA) which is concerned with the processes involved in acquiring a second language, including those of language transfer and cross-linguistic influence. Studies conducted in the field of SLA may be classified into different groups (*e.g.* qualitative or quantitative) based on the applied methodology. One such taxonomy of SLA research, put forward by Lightbown (1985), proposes the following three broad categories:

Descriptive studies which involve the examination of gathered linguistic data (*e.g.* learner data), comparing it against some other data (*e.g.* native language) and describing the differences and consistencies therein. Such studies are generally exploratory in nature. Once the "raw" data has been described and explored, researchers may also begin interpreting their initial findings and forming hypotheses about the observations.

Hypothesis-testing studies start with some hypothesis about learner's use of language and seek to test this, instead of examining natural language samples. Some of these hypotheses may have been formulated using results from the above-mentioned descriptive studies or alternatively, they may also come from other sources such as a Contrastive Analysis between two or more languages. This approach may then include the collection of additional data (*e.g.* through elicitation or testing) and most commonly involve the application of statistical significance testing (Norris, 2015).

Experimental studies are those which attempt to manipulate some variable, *e.g.* those relating to language instruction, and attempt to measure related outcomes. Such studies also often test a specific hypothesis; they are also more difficult to design and conduct.

Based on this taxonomy we can see that hypothesis generation plays an important role in SLA and can guide the direction of experimental work; exploratory studies help formulate hypotheses which are then confirmed or rejected by additional, potentially empirical, follow-up studies.

In exploratory studies, SLA research aims to find distributional differences in language use and error rates between L1s, often referred to as **overuse**, the extensive use of some linguistic structures, and **underuse**, the underutilization of particular structures, also known as *avoidance* (Gass and Selinker, 2008). These concepts relate to language transfer effects, and were described earlier in §2.4.2.

However, the identification of such linguistic patterns that could form the basis of a hypothesis — *e.g.* specific types of errors, idiosyncratic productions and their contexts — is a process which often involves substantial effort. We now look at an example of an SLA study to better understand what constitutes such a hypothesis and the steps involved in testing it.

An example of SLA research that examines a specific hypothesis is the study of Diéz-Bedmar and Papp (2008), comparing Chinese and Spanish learners of English with respect to the English article system (*a*, *an*, *the*). Informed by prior descriptive and empirical research and a Contrastive Analysis of the languages, they form specific hypotheses that they seek to test:

> Our prediction is that, since the Chinese language lacks a grammaticalised article system, Chinese learners of English face a grammatical learning task which will manifest itself in their underuse of the English articles due to the omission of the and *a*/*an* and a higher rate of use (overuse) of the zero (∅) article. On the other hand, [...] we expect that both

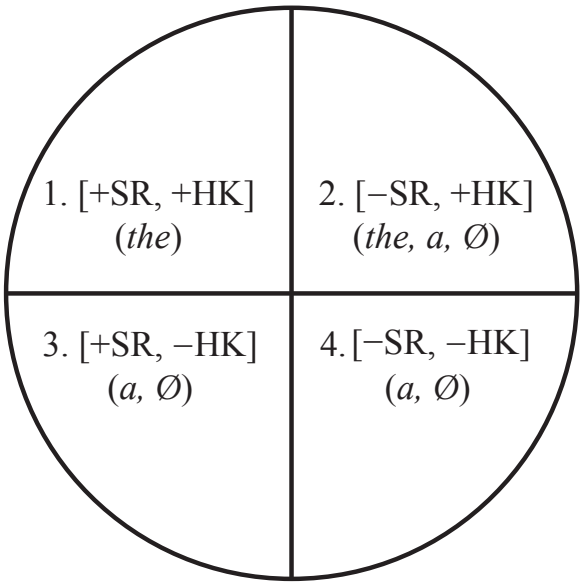

Figure 6.1: An illustration of Bickerton's semantic wheel showing the features associated with each category as well as the applicable articles.

> Chinese and Spanish learners will exhibit a pragmatic problem by misusing and overusing articles. While for the Spanish speakers this prediction means that we expect them to misuse and overuse English articles in their interlanguage, for the Chinese these seemingly contradictory predictions mean that we would expect them not only to underuse, but also to misuse and overuse articles until they learn to restrict their use to the specific semantic contexts and pragmatic functions in which they are employed in English.

They adopt a particular theoretical framework, using Bickerton's (1981) semantic wheel for noun phrase reference for the analysis of their data. This semantic wheel is a taxonomy of English article usage based on two binary features: [± Specific Reference] or [± SR] and [± Hearer Knowledge] or [± HK]. Specificity is the speaker's intention to refer and Hearer Knowledge is the speaker's assumption about the familiarity of the hearer with the referent and their ability to infer it. The combinations of these two features result in four categories, as shown in Figure 6.1. These four categories can be termed as:

Category	**Examples**
(1) Referential definites	Pass me the pen. The first person to walk on the moon...
(2) Generics	Elephants have trunks. Fruit flourishes in the valley.
(3) First mentions and referential indefinites	A man phoned. I've bought a new car.
(4) Non-referential and non-specific indefinites	Alice is an accountant. I guess I should buy a new car.

Table 6.1: The four categories in Bickerton's semantic wheel.

The authors then collected corpora containing 175 non-native and 164 native texts. Articles in the native corpora were tagged using a coding system derived from the four categories of Bickerton's

semantic wheel. This annotation allows for a corpus-based Contrastive Analysis of the three L1s, which the authors use as a basis of forming their hypotheses. The learner texts were also corrected by native speakers and incorrect article use was tagged with corrections, using the same coding system as the native texts.

Next, they use the simple Wordsmith software package — designed to extract data for lexicographers — to retrieve and quantify the tagged errors in a semi-automatic way, and evaluate whether Chinese and Spanish L1 speakers do behave differently via hypothesis testing (chi-square and z-tests, in their case). They conclude that Chinese and Spanish do have characteristic differences, with patterns of zero article and definite article use differing according to semantic context. Such studies are typically carried out on relatively small datasets, and use fairly elementary tools. Sources such as Ellis (2008) and Ortega (2009) give good overviews of such studies.

There are several aspects of such research, *e.g.* annotation and detection of overuse/underuse trends, where NLP and machine learning methods could help. These computational techniques hold out the promise of aiding these processes in a (semi-)automated way, enabling them to be conducted on large-scale data of a magnitude previously unprecedented in SLA research. This, in turn, could enable the identification of previously hidden issues in non-native speaker writing, and allow hypotheses to be tested on a much larger scale and with more empirical reliability than previously possibly. The remainder of this chapter is dedicated to our initial investigations into the plausibility of such an approach.

6.2 Related Work

6.2.1 NLI and Feature Analysis

While Native Language Identification (NLI) as a subfield of NLP has seen much new work in the last few years — §2.3.3 provides a comprehensive review — the emphasis on optimizing classification task results, for example by using classifier ensembles as we showed in Chapter 3, versus analysing features for relevance to other tasks has varied. In this section we briefly discuss works which directly look at how features might be related to language-learning tasks or SLA research.

The seminal work of Koppel et al. (2005a) that presented NLI as a classification task included, in addition to standard lexical and PoS n-gram features, errors made by the writers; these errors were automatically identified using Microsoft Word grammar checker. Kochmar (2011) used the First Certificate in English (FCE) corpus for NLI, including the manually annotated errors as features, and presented an analysis of usefulness of features (including errors) with respect to L1.

Wong and Dras (2011) used syntactic features on the basis of SLA theory that posits that L1 constructions may be reflected in some form of characteristic errors or patterns in L2 constructions to some extent, or through overuse or avoidance of particular constructions in L2 (Lado, 1957; Ellis, 2008); they did note distributional differences of features related to L1. Wong et al. (2012) induced topic models over function words and POS n-grams, where some of the topics appeared to reflect L1-specific characteristics. These works, while interested in the nature of the features, do not evaluate them except via classification accuracy. Coming from a linguistic perspective, the work in Jarvis and Crossley (2012) uses Linear Discriminant Analysis for classification of texts by L1, and identify interesting features by a stepwise feature selection process in the course of classification, rather than via the measurement of their variability across L1s. Swanson and Charniak (2012) similarly explore

using syntax, where they propose a richer representation for L1-specific constructions through Tree Substitution Grammar fragments, which we described in §2.3.2.8.

Among this efflorescence of NLI work, a new trend explored by researchers aims to extract lists of candidate language transfer features. Swanson and Charniak (2013) examine both relevancy and redundancy of features through a number of metrics (including the χ^2-statistic used later in this chapter). They then extend a Bayesian induction model for TSG inference based on a supervised mixture of hierarchical grammars, in order to extract a filtered set of more linguistically informed features that could benefit both NLI and SLA research; an aim was to find relatively rare features that are nevertheless useful for L1 prediction.

Swanson and Charniak (2014) continue on from this with a data-driven approach to ranking features, again using TSGs. They do this by comparing L2 data against the writer's L1 to find features where the L1 use is mirrored in L2 use. A key shortcoming here is that only the most obvious effects can be detected, but this is not the case with many transfer effects that are "too complex" to observe in this manner (Jarvis and Crossley, 2012, p. 183). Moreover, this method is unable to detect underuse, it is only suitable for syntactic features since it relies on constituent tree structures. It has also only been applied to very small data (4,000 sentences) over 3 L1s.

6.2.2 Relation to Language Teaching and Learning

The large demand for result-oriented language teaching and learning resources is an important motivating factor in SLA research (Richards and Rodgers, 2014; Ortega, 2009). Today, we live in a world where there are more bilingual individuals than monolinguals, but multilingualism does not automatically imply having attained full mastery of multiple languages. As the world continues on the path to becoming a highly globalized and interconnected community, the learning of foreign languages is becoming increasingly common and is driven by a demand for language skills (Tinsley, 2013). All of this provides intrinsic motivation for many of the learners to continue improving their language skills beyond that of basic communication or working proficiency towards near-native levels. In itself, this is not an easy task, but a good starting point is to reduce those idiosyncratic language use patterns caused by the influence of the native language. The first step towards this is to identify such usage patterns and transfer effects through studies such as this one.

The motivations for identifying L1-related language production patterns are manifold. Such techniques can help SLA researchers identify important L1-specific learning and teaching issues. In turn, the identification of such issues can enable researchers to develop pedagogical material that takes into consideration a learner's L1 and addresses them. This equates to teaching material that is tailored for students of an L1 group. Some research into the inclusion of L1 knowledge in teaching material has already been conducted.

Horst et al. (2010) investigated how L1 knowledge can be incorporated into language instruction in order to facilitate learning. They approached this by designing a series of cross-linguistic awareness (CLA) activities which were tested with francophone learners of English at a school in Montreal, Quebec, Canada. The CLA material was developed by identifying commonalities between French and English. Next, a set of 11 CLA teaching packages were developed and piloted in an intensive year-long ESL program. Although they did not conduct empirical evaluation with a control group, observations and interviews indicate that this is a promising approach that can address a wide range of linguistic phenomena.

Laufer and Girsai (2008) investigated the effects of explicit contrastive analysis on vocabulary acquisition. Three groups of L2 English learners of the same L1 were used to form separate instructional conditions: meaning focused instruction (MFI), non-contrastive form-focused instruction (FFI), and contrastive analysis and translation (CAT). The CAT group performed translation tasks and was also provided a contrastive analysis of the target items and their L1 translation options. One week later, all groups were tested for retention of the target items and the CAT group significantly outperformed the others. These results are interpreted as evidence for L1 influence on L2 vocabulary acquisition.

Such findings from SLA research, although not the principal source of knowledge for teachers, are considered helpful to them and have great pedagogical relevance. Although a more comprehensive exposition of the pedagogical aspects of SLA is beyond the scope of our work, we refer the interested reader to Lightbown (2000) for an overview of SLA research in the classroom and how it can influence teaching.

6.3 Extracting Language Transfer Hypothesis Candidates

As we described earlier, SLA researchers are interested in identifying distributional differences that could be attributed to language transfer effects — *i.e.* overuse and underuse — in learner language. While there have been some attempts in SLA to use computational approaches on small-scale data,[1] these still use fairly elementary techniques and have several shortcomings, including in the manual approaches to annotation and the computational artefacts derived from these.

Conversely, NLI work has focused on automatic learner L1 classification using machine mearning with large-scale data and sophisticated linguistic features. Our NLI experiments from the preceding chapters have demonstrated that the linguistic productions of different L1 groups can be distinguished with high accuracy. NLI experiments by other researchers, as discussed in §2.3.3, have also yielded good results.

However, despite these well-established results in predicting the L1s of non-native authors, few attempts have been made to interpret the features that distinguish the L1s. This is partly because no methods for an SLA-oriented feature analysis have been proposed; most work has focused on testing feature types using standard machine learning tools.

The overarching aim of this section is to develop a methodology that enables the exploration of discriminative NLI features to examine their use for SLA research. This can help us evaluate if this is a plausible approach for connecting the NLI paradigm to SLA applications that could be used to link these features to their underlying linguistic causes and explanations. These candidates can then be applied in other areas such as remedial SLA strategies.

We propose that feature weights calculated in training a learning algorithm are a potentially interesting way to explore the feature space and argue that they are more suitable that other ranking and relevancy methods used for feature selection. In the next section we describe the proposed method, followed by an outline of our data and feature types, and we conclude with the results of our investigation in the final section.

[1] *e.g.* Chen (2013) explore the use of phrasal verbs by Chinese students using approximately 1800 essays. Similarly, Lozanó and Mendikoetxea (2010) investigated the production of postverbal subjects (*i.e.* VS order) in L2 English, using some 250k words of non-native data. Such studies are on a much smaller scale compared to the investigation we present using 4m tokens across 12k essays from TOEFL11.

6.3.1 Language Transfer Hypothesis Candidate Extraction Method

In order to extract features which are strongly associated with an L1 group, *i.e.* the overuse and underuse features described above in §6.1, some type of feature selection method (Liu and Motoda, 2007) could be applied to extract the most informative features. The feature sets obtained via such methods can then be "analysed by domain experts to gain more insight into the problem modelled" (Saeys et al., 2008) and better understand the process from which the data was extracted. One example of such a feature scoring method is the Fisher score:

$$F(j) \equiv \frac{\left(\bar{\boldsymbol{x}}_j^{(+)} - \bar{\boldsymbol{x}}_j\right)^2 + \left(\bar{\boldsymbol{x}}_j^{(-)} - \bar{\boldsymbol{x}}_j\right)^2}{\frac{1}{n_+ - 1}\sum_{i=1}^{n_+}\left(x_{i,j}^{(+)} - \bar{\boldsymbol{x}}_j^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{i=1}^{n_-}\left(x_{i,j}^{(-)} - \bar{\boldsymbol{x}}_j^{(-)}\right)^2} \tag{6.1}$$

The F-score (Fisher score, not to be confused with F-measure) measures the ratio between the intraclass and interclass variance in the values of feature j, where x represents the feature values in the negative and positive examples.[2] More discriminative features have higher scores.

Another alternative method is Information Gain (Yang and Pedersen, 1997). As defined in equation 6.2, it measures the entropy gain associated with feature t in assigning the class label c.

$$\begin{aligned} G(t) = & - \sum_{i=1}^{m} Pr(c_i) \log Pr(c_i) \\ & + Pr(t) \sum_{i=1}^{m} Pr(c_i|t) \log Pr(c_i|t) \\ & + Pr(\bar{t}) \sum_{i=1}^{m} Pr(c_i|\bar{t}) \log Pr(c_i|\bar{t}) \end{aligned} \tag{6.2}$$

Until the late 1990s, machine learning work in most domains used only tens of features (Guyon and Elisseeff, 2003), although this has rapidly expanded to datasets with thousands, if not millions, of features for complex tasks like gene selection or text classification. Although the abovementioned methods may yield reasonable results for small or trivial data with relatively few features, it has been noted that they fail to produce good results for large numbers of features in multi-class settings (Forman, 2004).

Other limitations of these methods, particularly related to our task, are that they do not provide ranked lists per class, and more importantly, they do not explicitly capture per class underuse, an important concept in SLA. In sum, the ability of these methods for inspecting features to extract useful knowledge from the data is limited.

Another approach that has been proposed to address this issue is based on using an SVM classifier to evaluate feature relevance and extract information from large data. In their innovative and influential work, Guyon et al. (2002) show that SVMs have both qualitative and quantitative advantages over other feature ranking and selection methods. They also highlight that a key advantage of SVMs is their ability to take into account the mutual information captured by the input features, whereas measures such as Information Gain or the F-score assess features independently. Furthermore, although many learning algorithms require feature selection for input pruning to reduce computational

[2] See Chang and Lin (2008) for more details.

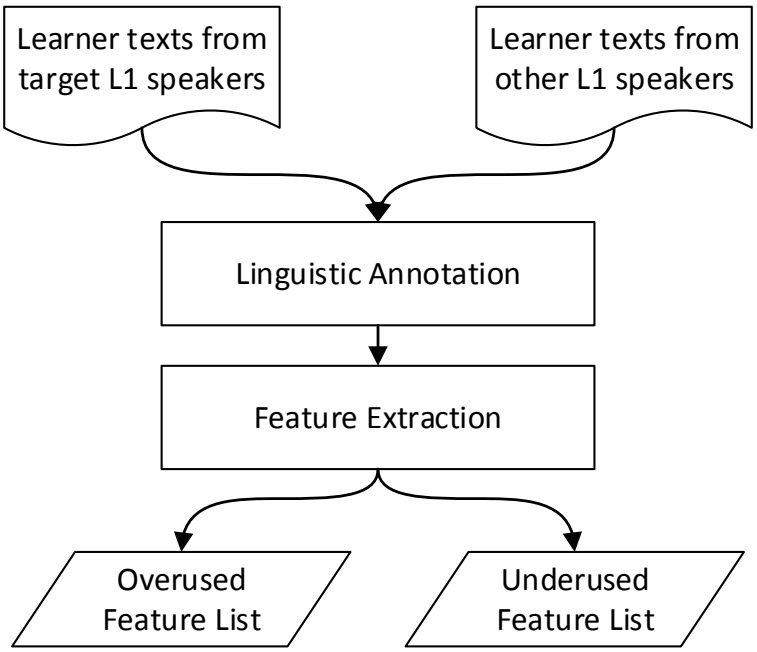

Figure 6.2: Our methodology for extracting language transfer hypothesis candidates is based on contrasting L1 groups against each other to find discriminative features.

cost, SVMs are known to be robust to very high dimensional input and feature selection is generally not needed (Rogati and Yang, 2002). In this manner, SVMs have been successfully applied in data mining and knowledge discovery tasks such as examining proteomic data for tumor diagnostics analysis (Jong et al., 2004) and identifying discriminant cancer genes (Guyon et al., 2002).

Our approach to identifying potential language transfer effects follows this methodology and is based on the use of SVM weights to identify distinguishing linguistic features of each L1. We take the very same classification approach as in previous chapters, described in §3.2.1. One aspect that bears repeating is that we adopt a one-vs-all (OVA) approach in adapting SVMs to multi-class data, as described in §3.1.3.1. Here, this also allows us to find features most relevant to each native language class. The classifier was trained using l_2 regularization. An overview of this approach is shown in Figure 6.2. Our data and features will be described in §6.3.2 below.

Using the extracted features, we train linear Support Vector Machine (SVM) models for each L1. In training each model, the SVM weight vector is calculated according to (6.3):

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \tag{6.3}$$

Here $y_i \in \{1, -1\}$ and represents the class label and $\mathbf{x}_i$ is the feature vector for instance i.[3] The weight vector $\mathbf{w}$ is calculated for each L1 class. Features associated with the positive class will have larger, positive weights and unrelated features will be assigned negative weights.

After training, the positive and negative weights for each L1 are split into two lists and ranked by weight. The positive weights represent overused features, while features whose absence (*i.e.* underuse) is indicative of an L1 will have large negative weights. This process yields two candidate language transfer feature lists per L1.

[3] See Burges (1998) for a detailed explanation.

6.3.2 Data and Features

We use the TOEFL11 corpus for this investigation, which was described in §2.3.1.4. The following feature types were used in our analysis of the learner texts. They were extracted using the same methodology we earlier described in §3.1.3.2.

Adaptor grammar collocations We use the mixed POS and function word version to discover arbitrary length n-grams, as described in §2.3.2.10.

Stanford Dependencies This feature is used to capture longer distance grammatical relations and their types, as described in §2.3.2.9.

Lexical features Content and function words are also considered as two feature types related to learner's vocabulary and spelling. See §2.3.2.3.

6.3.3 Results

We now turn to an analysis of the output from our system to see how it might be useful for SLA research. Table 6.2 lists some elements from the underuse and overuse lists for various L1s. The lists are of different feature types. They have been chosen to demonstrate all feature types, but also a wide range of languages. We look at some examples of highly weighted features that have been discussed in the SLA literature.

Hindi L1 writers are distinguished by certain function words including *hence*, *thus*, and *etc*, and a much higher usage rate of male pronouns. It has been observed in the literature (Sanyal, 2007, for example) that the English spoken in India still retains characteristics of the English that was spoken during the time of the Raj and the East India Company that have disappeared from other English varieties, so it sounds more formal to other speakers, or retains traces of an archaic business correspondence style; the features noted fit that pattern.

The second list includes content words overused by Arabic L1 learners. Analysis of content words, and for many other L1s in our data, reveals very frequent misspellings which are believed to be due to orthographic or phonetic influences (Tsur and Rappoport, 2007; Odlin, 1989). Since Arabic does not share orthography with English, it is possible that most of these are due to phonetics. Looking at items 1, 3 and 5 we can see a common pattern: the English letter *u* which has various phonetic realizations is being replaced by a vowel that more often represents that sound. Items 2 and 5 are also phonetically similar to the intended words.

For Spanish L1 authors we provide both underuse and overuse lists of syntactic dependencies. The top 3 overuse rules show the word *that* is very often used as the subject of verbs.[4] This is almost certainly a consequence of the prominent syntactic role played by the Spanish word *que* which, depending on the context, is equivalent to the English words *whom*, *who*, *which*, and most commonly, *that*. The 4^{th} rule shows they often use *this* as a determiner for plural nouns. A survey of the corpus reveals many such errors in texts of Spanish learners, *e.g.* "*this actions*" or "*this emissions*". The 5^{th} rule shows that the adjectival modifier of a plural noun is being incorrectly pluralised to match the

[4] Although *that* is considered a Wh-determiner in this case, this is the relationship which the Stanford dependency parser assigns to such tokens in a relative clause. The relationship between the head of the preceding relative clause and the verb or noun is usually marked separately, as illustrated in the third example in Figure 6.3.

Overuse			**Underuse**	
Hindi	Arabic	Spanish	Spanish	Chinese
#2: thus	#2: anderstand	#1: `nsubj(VBP,that)`	#2: `det(NNS,these)`	#12: *an* `NN`
#4: hence	#4: mony	#2: `nsubj(VBZ,that)`	#3: `nsubj(VBZ,which)`	#16: *other* `NN`
#22: his	#6: besy	#3: `nsubj(VB,that)`	#6: `nsubj(VB,which)`	#18: *these* `NNS`
#30: etc	#15: diffrent	#4: `det(NNS,this)`	#7: `nsubj(VBP,which)`	#19: *even if*
#33: rather	#38: seccessful	#25: `amod(NNS,differents)`	#10: `det(NN,no)`	#68: *might*

Table 6.2: Example transfer features and ranks from the overuse/underuse lists for various L1s and features, in order: Hindi function words, Arabic words, Spanish dependencies and Chinese Adaptor Grammars.

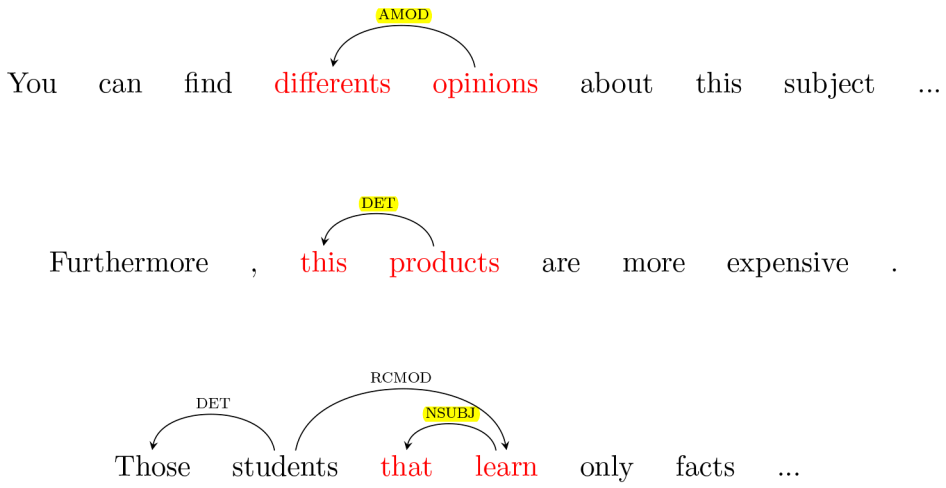

Figure 6.3: Three of the most common overuse patterns found in the writing of L1 Spanish learners. They show erroneous pluralization of adjectives, determiner misuse and overuse of the word *that*.

noun in number as would be required in Spanish, for example, "differents subjects". Some examples of these dependencies are shown in Figure 6.3.

Turning to the underused features in Spanish L1 texts, we see that 4 related features rank highly, showing that *these* is not commonly used as a determiner for plural nouns and *which* is rarely used as a subject. The final feature shows that *no* is avoided as a determiner. This may be because while *no* mostly has the same role in Spanish as it does in English, it cannot be used as a determiner; *ningún* must be used instead. We hypothesize that this construction is being avoided as placing *no* before a noun in Spanish is ungrammatical. This example demonstrates that our two list methodology can not only help identify overused structures, but also uncovers the related constructs that are being underutilized at their expense.

The final list in Table 6.2 is of underused Adaptor Grammar patterns by Chinese learners. The first two features show that these writers significantly underuse determiners, here *an*, *other* and *these* before nouns. This is not unexpected since Chinese learners' difficulties with English articles are well known (Robertson, 2000). More interestingly, we find underuse of features like *even if* and *might*, along with others not listed here such as "*could* `VB`"[5] plus many other variants related to the subjunctive mood. To illustrate this, Figure 6.4 shows the number of exam scripts in each L1 group that contain the phrase *even if*, highlighting the large inter-group differences in the use of this linguistic feature.

One explanation is that linguistic differences between Chinese and English in expressing counterfactuals could cause them to avoid such constructions in L2 English. Previous research in this area has linked the absence of subjunctive linguistic structures in Chinese to different cognitive representations of the world and difficulties in thinking counterfactually (Bloom, 2014), although this has been disputed by others (Au, 1983; Garbern Liu, 1985).

We now discuss several features relating to languages not listed in Table 6.2. Adaptor Grammars also reveal frequent use of the "existential *there*"[6] in German L1 data while they are highly underused

[5] *e.g. could be*, *could have*, *could go* and other variants
[6] *e.g.* "There is/are ...", different to the locative "there"

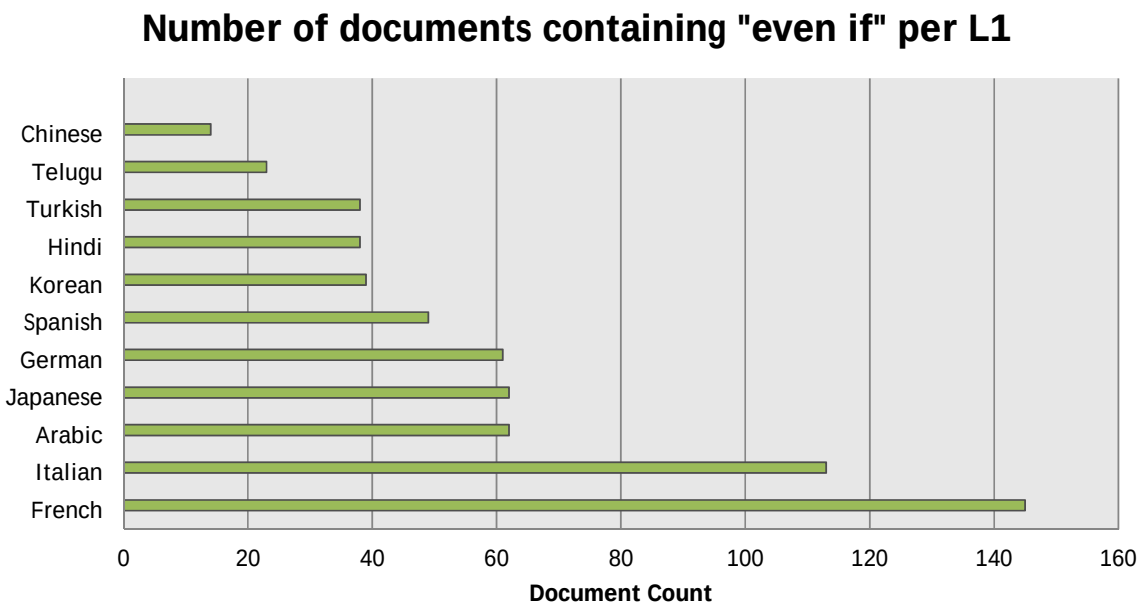

Figure 6.4: The number of exam scripts in each L1 group of the TOEFL11 corpus containing the phrase "even if" in the text. The total number of documents per L1 are equal.

English	**Spanish**	**English**	**Spanish**
diferent	diferente	conclution	conclusión
consecuence	consecuencia	desagree	Neg. affix *des-*
responsability	responsabilidad	especific	específico
oportunity	oportunidad	necesary	necesario

Table 6.3: Common and highly ranked English misspellings of Spanish learners and their Spanish cognates.

in French L1 data. The literature is in agreement with our data: The German equivalent *es gibt* is common while French use is far more constrained (Cappelle and Loock, 2013). The use of the existential *there* in French was studied by Cappelle and Loock (2013), who state:

> Drawing on extensive counts conducted in available corpora and self-compiled samples of translated English and French, intra-language comparisons of translated and non-translated language use show that existential *there* is under-represented in English translated from French while existential *il y a* is over-represented in French translated from English. It is suggested that source-language interference is responsible for these differences.

Lexical analysis also revealed Spanish–English orthographic transfer in the texts of L1 Spanish authors, with some highly ranked patterns listed in Table 6.3. This transfer is evidenced by the similarity of misspelled English words to their Spanish equivalents. We observe transfer of orthographic rules and conventions (*e.g.* omission of double letters in *diferent* and *oportunity*) as well as morphemes (*e.g. desagree* with the Spanish affix *des-* being equivalent to *dis-* in English).

We also observe other patterns which remain unexplained. For instance, Chinese, Japanese and Korean speakers make excessive use of phrases such as *however*, *first* and *second*. One possibility is that this relates to argumentation styles that are possibly influenced by cultural norms. More broadly,

this effect could also be teaching- rather than transfer-related. For example, it may be case that a widely-used text book for learning English in Korea happens to overuse this construction. While the overuse is extreme in some cases,[7] more research is required to determine the causal factors. We hope to expand on this and other hypotheses in future work.

Some recent findings from the 2013 NLI Shared Task found that L1 Hindi and Telugu learners of English had similar transfer effects and their writings were difficult to distinguish. It has been posited that this is likely due to shared culture and teaching environments (Malmasi et al., 2013).

6.3.4 Discussion

In this section we defined a method for ranking arbitrary features that reflect both overuse and underuse for individual L1s. We believe that these results highlight some of the interesting hypotheses and patterns that can be extracted from learner data using our methodology. More specifically, this investigation shows that the feature weights assigned by the SVM have some correspondence with phenomena and hypotheses discussed in the SLA literature; this evidences the suitability of our method for identifying features that can form the basis of SLA hypotheses as we discussed in §6.1. This investigation, an intersection of NLP, Machine Learning and SLA, illustrates how the various disciplines can complement each other by bringing together theoretical, experimental and computational issues.

In terms of NLI, this work is the first attempt to examine and present a broad linguistic interpretation of the features that enable learning algorithms to accurately classify L1s, including feature underuse. NLI systems achieve classification accuracies of over 80% on this 11-class task, leading to theoretical questions about the features that make them so effective. This work also has a backwards link in this regard by providing qualitative evidence about the underpinning linguistic theories that make NLI work.

There are several directions for future work. The first relates to clustering the data within the lists. Our intuition is that there might be coherent clusters of related features, with these clusters characterising typical errors or idiosyncrasies, that are predictive of a particular L1. As shown in our results, some features are highly related and may be caused by the same underlying transfer phenomena. For example, our list of overused syntactic constructs by Spanish learners includes 3 high ranking features related to the same transfer effect. The use of unsupervised learning methods such as Bayesian mixture models may be appropriate here. For parse features, tree kernels could help measure similarity between the trees and fragments. Finally, the use of other linguistic features such as Context-free Grammar phrase structure rules or Tree Substitution Grammars could provide additional insights.

An important extension of this work could entail the inclusion of native speaker data in the experiment. The current experiment was performed using non-native groups and focused on identifying the differences between these groups. Previously, in §5.7, we demonstrated that native and non-native text could be distinguished with great accuracy. However, it is not clear how the addition of a native speaker class to the standard NLI performance would affect its performance. For identifying language transfer effects, such a setup could help identify at least two things: (1) the set of idiosyncratic constructs common to the majority of learners; and (2) constructions and vocabulary not acquired by any of the non-natives. Both of these would be of high value to researchers. The main challenge

[7] Over half the Korean texts had at least one sentence-initial *however*.

to performing such experiments is the paucity of native speaker data covering the same topics and collected under the same conditions, as we pointed out in §5.7.1.

Another important aspect to consider is the introduction of appropriate evaluation methods. Given the recency of this topic, no standard evaluation measures have been proposed for this emergent area of research. Accordingly, a key goal of future work should be the development of such evaluation methods for this nascent task.

The evaluation of the extracted hypotheses is based on the interpretation of SLA experts and consequently, quantitative evaluation is not possible *per se*, without expert evaluation. Therefore, it is not possibly to quantitatively evaluate the quality of the lists produced by each method, this requires human evaluation and thus the features produced through such methodology are hypothesis *candidates*.

It is however possible to develop similarity metrics and methods for comparing the candidate lists and this is an issue we plan to address in future research. This enables the direct comparison of candidate lists produced by different methods, researchers and any previously proposed methods. It also allows researchers to compare lists from different L1 groups to identify similar patterns of language use that may be shared across specific groups of L1s. We also aim to investigate the development of additional metrics to assist experts in interpreting the results. This includes the relevant statistical analyses as well as visualization methods.

Notwithstanding these directions, in the next section we go on to look at how to use these sorts of features to construct and rank SLA hypotheses.

6.4 Extracting Error Contexts

As noted in §6.1, a research goal in SLA is to formulate and test hypotheses about errors and the particular environments in which they are made. There are several steps that would need to be taken to use the information from Eqn. 6.3 to formulate an SLA hypothesis. Typically hypotheses involve more than just a single (type of) word.

One particular approach to finding aspects of texts characteristic of their L1s that has motivated the work in this section is described in Yannakoudakis et al. (2012), the goal of which is to develop visualisation tools for SLA researchers. They present graphs of the relationships between errors and their contexts, such that SLA researchers can navigate through the graphs to find contexts for particular errors that can lead to hypotheses like that of Diéz-Bedmar and Papp (2008) above. In this section, we look at approaches to finding such hypothesis candidates automatically in the context of L1–L2 interaction by analysing the graphs used in the visualisations of Yannakoudakis et al. (2012). Specifically, we do the following:

- We propose a new task that is more directly oriented to SLA research than NLI has been for the most part, with the goal of identifying error-related contexts that are characteristic of L1s.
- We evaluate a number of models for finding such contexts, ranging from a simple baseline to treating the problem as a graph-theoretic maximum weighted clique one.
- We examine the results of some of the models to see how the task and the models might contribute to SLA research.

Verb Agreement	Some people <ns type="AGV">says<c>say</c></ns> ...
Incorrect Verb Inflection	The day I <ns type="IV">shaked<c>shook</c></ns> their hands,...
Missing Determiner	I am <ns type="MD"><c>a</c></ns> really good singer.

Figure 6.5: FCE corpus examples. Error types indicated by <ns type>...</ns>; errors indicated by ...; corrections indicated by <c>...</c>.

We began this chapter with some details on related work in §6.2. Although that body of work shares some similarities with the present experiment, our focus is on errors rather than on the distributional differences, and we look at error contexts that may not constitute a TSG tree or grammatical dependency.

Because we draw heavily on the work of Yannakoudakis et al. (2012), we next review relevant aspects of that work in §6.4.1; we then present our task definition and experimental setup in §6.4.2; we give results along with a discussion in §6.4.3; and we conclude in §6.4.4.

6.4.1 Developing Hypotheses: A Visualisation Tool

The context of the Yannakoudakis et al. (2012) work is automated grading of English as a Second or Other Language (ESOL) exam scripts, as described in Briscoe et al. (2010). The automated grading takes a classification approach, using a binary discriminative learner, with useful features including lexical and POS n-grams.

The publicly available dataset on which the work was carried out consists of texts from the FCE exam, aimed at upper-intermediate students of English across various L1s, and was presented in Yannakoudakis et al. (2011). This FCE corpus[8] consists of a subset of 1244 texts of the Cambridge Learner Corpus,[9] and is manually annotated with errors and their corrections, as well as a classification according to an error typology, as in Figure 6.5.

Yannakoudakis et al. (2012) present their English Profile (EP) visualiser as a way to "visually analyse as well as perform a linguistic interpretation of discriminative features that characterise learner English", using the features of this essay classification task. A screenshot of the front-end for their system in shown in Figure 6.6.

They define a measure of co-occurrence of features, among themselves and with errors, as a core part of their analysis. Given the set of all sentences in the corpus $S = \{s_1, s_2, \ldots, s_{|S|}\}$ and the set of all features $F = \{f_1, f_2, \ldots, f_{|F|}\}$, a feature $f_i \in F$ is associated with a feature $f_j \in F$ $(i \neq j, 1 \leq i, j \leq M)$ according to the score given in Equation (6.4), for $s_k \in S, 1 \leq k \leq N$ and exists() a binary function returning true if the input feature occurs in s_k.

$$\text{score}_{\text{ff}}(f_j, f_i) = \frac{\sum_{k=1}^{|S|} \text{exists}(f_j, f_i, s_k)}{\sum_{k=1}^{|S|} \text{exists}(f_i, s_k)} \tag{6.4}$$

[8] http://ilexir.co.uk/applications/ep-visualiser/
[9] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item364603/

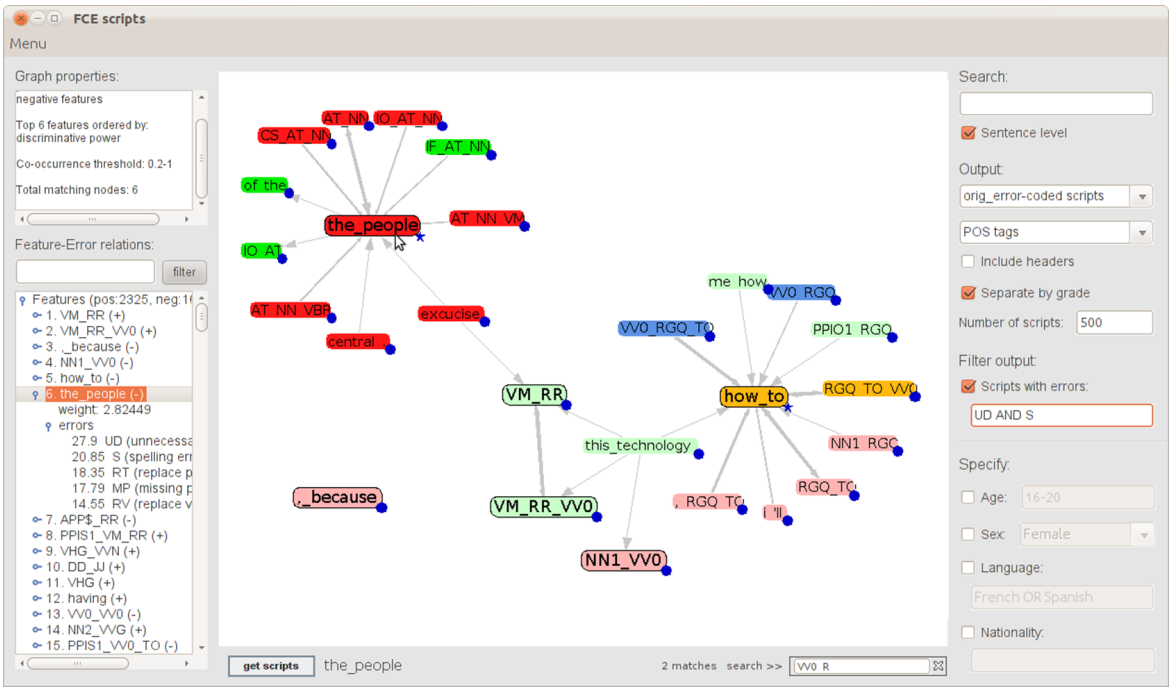

Figure 6.6: The front-end of the EP visualizer system. Image reproduced from Yannakoudakis et al. (2012).

They mention an analogous measure for feature-error co-occurrence; we assume given the set of all errors $E = \{e_1, e_2, \ldots, e_{|E|}\}$ that this is defined as follows:

$$\text{score}_{\text{ef}}(f_j, e_i) = \frac{\sum_{k=1}^{|S|} \text{exists}(f_j, e_i, s_k)}{\sum_{k=1}^{|S|} \text{exists}(e_i, s_k)} \tag{6.5}$$

A graph is defined with features and errors as vertices; an edge between features (resp. features and errors) is established if $\text{score}_{\text{ff}}()$ (resp. score_{ef}) is within some user-defined range. This graph of feature–feature (resp. feature–error) relationships is then presented visually. An example of such a feature-feature graph is shown in Figure 6.7.

The paper then presents a case study of how the EP visualiser can be used to assist SLA researchers. The case study starts by noting that RG_JJ_NN1 is the 18th most discriminative negative feature from the essay classifier; then, further inspecting the graph of discriminative features, that it's linked to JJ_NN1_II and VBZ_RG. Then, looking at feature-error relations, it investigates an association with error MD (missing determiner), and presents some examples that match the features (*e.g. Unix is very powerful system but there is one thing against it*), along with a discussion of relationships to various L1s. It is this process of finding interesting features and linking them to particular errors and L1s that we present an approach to automating in this section.

6.4.2 Task Definition and Experimental Setup

At a general level, our goal is to find which kinds of constructions (in a loose sense) centred around errors are particularly characteristic of various L1s.

The specific task we define for this paper, then, is to select a set of features (in the terminology of Yannakoudakis et al. (2012))—which we refer to as the ERROR CONTEXT—that, when combined with the error, show a strong association with L1, in a manner we describe below. So, for example,

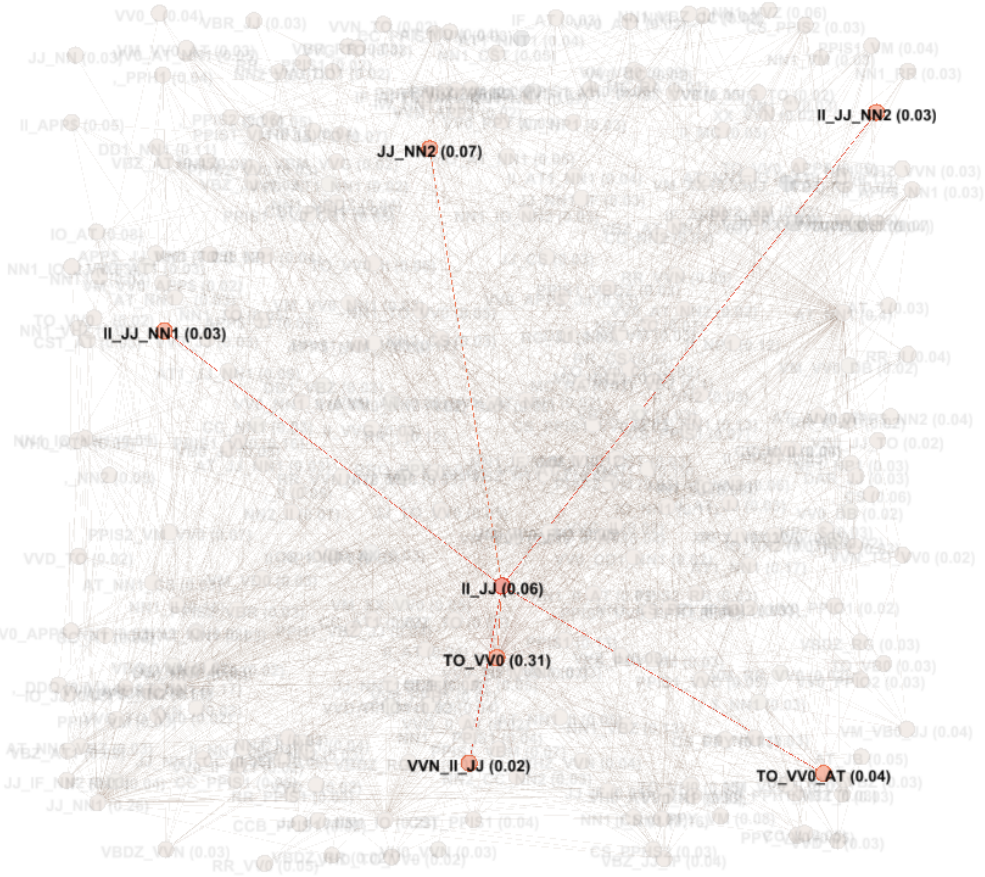

Figure 6.7: An example of a feature-feature graph for the feature II_JJ.

this may involve finding that an MD error in the context of RG_JJ_NN1, JJ_NN1_II and VBZ_RG shows a strong association with L1. We investigate a number of models for this selection process: the task then is the identification of which models produce poor error contexts (which will not rank highly in hypothesis testing) and which produce good ones (potentially worth considering by an SLA researcher). Below we discuss the data we use, the measure of association for an error and its context, the set of errors chosen, and the models for selecting context.

Language	Size
Chinese CHI	66
French FRE	146
German GER	69
Italian ITA	76
Japanese JAP	81
Korean KOR	86
Spanish SPA	200
Turkish TUR	75

Table 6.4: FCESUB, broken down by language

language	mean
CHI	0.885790
FRE	0.460894
GER	0.366587
ITA	0.581401
JAP	1.058159
KOR	1.067211
SPA	0.472253
TUR	1.014129
F-stat	18.031
sig.	<0.001

Table 6.5: ANOVA results giving mean score (number of sentences with MD error per 10 sentences) for each language, the ANOVA F-statistic, and significance value

6.4.2.1 Data

The corpus we use for evaluating the models for our task is derived from the FCE corpus of Yannakoudakis et al. (2012). The full FCE corpus consists of 1244 scripts over 16 languages; script counts range from 2 (Dutch) to 200 (Spanish).

The features used by Yannakoudakis et al. (2012) were derived from their essay classification task. As we are interested in associations with L1, we instead use features from the system we submitted to the NLI shared task and described in Chapter 3, which was applied to the TOEFL11 dataset. We only use POS n-grams ($n = 1, 2, 3$) as features in this section, as we had earlier found them to be good features.[10] Note that we use the terminology of Yannakoudakis et al. (2012) here: what had their origin as features in the essay classification task are still referred to as features in the visualisation tool, although the task carried out there is not a classification one. Similarly, we refer to our PoS n-grams as features, although we are not classifying errors using these features and so are not carrying out feature selection for the typical purpose of optimising classification performance.

For this, as did Yannakoudakis et al. (2012), we use the RASP parser (Briscoe et al., 2006) for tagging; the tags are consequently from the CLAWS2 tagset,[11] which are more fine-grained in terms of linguistic analysis than the more frequently used Penn Treebank tags.

For our task, we then used the subset of the FCE corpus where the languages overlapped with the TOEFL11 corpus: we refer to this as FCESUB. This gives 799 scripts over 8 languages, distributed as in Table 6.4; a positive byproduct is that the L1s are more similar in size than the full FCE corpus.

6.4.2.2 Association Measure

We noted in §6.1 that SLA studies such as Diéz-Bedmar and Papp (2008) use standard hypothesis testing techniques. We take this as a starting point. We could, for example, evaluate whether a particular raw error (that is, without a feature context) is strongly associated with L1s by using a single factor ANOVA test.[12] The independent variable would be the L1. The dependent variable could be one of a number of alternatives; we choose the number of sentences with a particular error

[10] See §3.1.3.5 and §5.8 for more details.
[11] `http://ucrel.lancs.ac.uk/claws2tags.html`
[12] See, *e.g.*, Jackson (2009).

Type	Name	F-stat	p-val	N
DJ	Wrong Derived Adjective	3.27	.002	332
DN	Wrong Derived Noun	0.70	.671	294
MD	Missing Determiner	18.03	.000	1702
MT	Missing Preposition	2.81	.007	985
UD	Unnecessary Determiner	1.20	.301	807
UT	Unnecessary Preposition	0.26	.968	689
UV	Unnecessary Verb	0.78	.606	317

Table 6.6: Error types chosen for evaluation, including F-statistic, ANOVA p-value and corpus count of sentences containing error.

per 10 sentences.[13] To illustrate, we give the ANOVA results from FCEsub for the MD error in Table 6.5. The ANOVA calculation is based on an F-statistic which compares variance between treatments against variance within treatments; this is compared against critical values for the F-statistic to determine statistical significance. The expected value of the F-statistic under the null hypothesis is 1, with values above 1 increasingly inconsistent with the null hypothesis. The data in Table 6.5 shows that the MD error does vary significantly with L1; a post-hoc Tukey HSD test lets us identify which specific languages exhibit this difference and shows that, for example (and as can be observed in the means), German L1 speakers are significantly different from Korean L1 speakers in the occurrence of MD errors.

For our task we are not interested in significance per se. Rather, we are interested in whether we can find occurrences of errors plus contexts that are more strongly associated with, or that vary across, L1s, *e.g.* that an MD error in the context of RG_JJ_NN1, JJ_NN1_II and VBZ_RG is more strongly associated with L1s; and we are also interested in which of our proposed methods for identifying an error's feature context does this best. For this purpose, then, we use just the F-statistic from the ANOVA test, this time with the dependent variable as the ratio of occurrences of error plus error context per 10 sentences: a higher F-statistic shows a stronger association with L1s.[14]

We also consider the χ^2-statistic from Pearson's chi-squared test, noting that it is also used in SLA hypothesis testing and that it was additionally found by Swanson and Charniak (2013) to be good at distinguishing interesting features in their related task (see §6.2 for more detail). The F-statistic and χ^2-statistic are closely related: a random variate of the F-distribution is the ratio of two chi-squared variates scaled by their degrees of freedom. A difference is that χ^2 compares observed versus expected counts rather than proportions: to take account of the differing text lengths, our observed frequency is the number of sentences with error and error context per L1; our expected frequency is the total number of sentences with that error and error context scaled according to the proportion of sentences labelled with that L1 relative to the corpus as a whole.

6.4.2.3 Errors Chosen

From the 74 error types in the FCE corpus, we select a subset to evaluate our models. In addition to the MD error used in the case study of Yannakoudakis et al. (2012), we choose a subset which has a range of F-statistic values as described above: some show very similar patterns across L1s (*i.e.* with

[13] We note that the texts differ significantly in length by L1, so it would not be suitable to normalise as occurrences per document.

[14] As we are only using the F-statistic to evaluate ranks, we do not need a multiple comparison adjustment such as the Bonferroni correction: this would only apply for comparisons to a significance threshold, and in any case the Bonferroni is monotonic and does not affect rankings.

low F-statistic), such as DN Wrong Derived Noun (e.g. *hot* vs *heat*); others do vary significantly with L1, such as DJ Wrong Derived Adjective (e.g. *reasonally* vs *reasonable*). Having errors with a range of F-statistic values lets us evaluate whether finding good error contexts works only for strongly L1-associated errors, weakly L1-associated errors, or across the spectrum. Our subset is in Table 6.6, along with their F-statistic, ANOVA p-value and counts in FCEsub.

6.4.2.4 Models

We propose four models for choosing error contexts. These models rank error contexts; we evaluate the ranked error contexts by F-statistic and χ^2-statistic values (§6.4.2.2).

ErrorCoocc In this model we rank features by error-feature co-occurrence scores given by Equation (6.5). The L1 is not taken into account, so this will just return common features which may be equally strongly associated with errors across all L1s. We look at results for when $k = 1..3$ features are chosen. For $k = 2, 3$, we add the individual error-feature scores together for the ranking.[15] It may be the case that interesting results could be obtained for $k > 3$, but we only look at the $k = 1..3$ in this preliminary work to see if there are any discernible trends suggesting that larger values of k could help.

L1Assoc Here we use features that are strongly associated with the L1s from the Toefl11 corpus and NLI shared task. Specifically, we rank features by their Information Gain with respect to L1s as in the process of feature selection from the shared task.[16] The relationship between errors and features (in the form of error-feature co-occurrence scores) is not taken into account here. Again, we look at results for when $k = 1..3$ features are chosen, and for $k = 2, 3$, we add the individual error-feature scores together for the ranking.

MaxWeightClique Both of the preceding models look only at one factor that might be relevant: error-feature scores (finding features that are related to the errors) and a measure of the association of features with L1s; but there is no link between them, and interaction of features is not taken into account. In Yannakoudakis et al. (2012), the visualiser provides to the SLA researcher a graph showing the relatedness of features, based on Equation (6.4), and the SLA researcher combines this with error-feature scores to find interesting candidate error contexts; we create a similar graph and aim to imitate the process by incorporating error-feature scores as follows.

We define a weighted undirected graph $G = (V, A)$ such that V is the set of features used in the above models (*i.e.* POS n-grams from ErrorCoocc); A is defined such that $(v_i, v_j) \in A$ for vertices $v_i, v_j \in V$ if $0.8 \leq \text{score}_{\text{ff}}(v_i, v_j) \leq 1.0$ where $\text{score}_{\text{ff}}()$ is as defined as in Equation (6.4).[17] Given our set of errors E defined at Equation (6.5) above, the weight of a vertex v_i is defined as $\text{score}_{\text{ef}}(v_i, e_j)$ for some $e_j \in E$. Given this graph, it is possible to characterise the finding of related features with strong aggregate associations with errors as an instance of the maximum weight clique problem (Bomze et al., 1999). As the name suggests, this finds a clique of maximum weight, here the strongest aggregate feature–error association. While this is an NP-hard problem, there are quite efficient algorithms for solving it; we use one called *wclique*, proposed by Östergård (1999).[18]

[15] For $k = 2$ the combinations were made from the top 100 features from $k = 1$, and for $k = 3$ from the top 50.
[16] We recalculated this over the subset of eight languages used in this paper.
[17] We choose this threshold value as it is the one used in the graph definition of Yannakoudakis et al. (2012).
[18] Code for *wclique* is available at: `http://tcs.legacy.ics.tkk.fi/~pat/wclique.html`.

model	r
ErrorCoocc	0.95
L1Assoc	0.97
MaxWeightClique	0.95
MaxWeightClique-L1	0.92

Table 6.7: Average correlation coefficient r between the F-statistic and chi-square statistic for each model

MaxWeightClique-L1 We also look at a variant of MaxWeightClique where we construct the graphs based only on relationships among features for a particular L1. That is, there will be eight weighted graphs per error of interest.

6.4.3 Results and Discussion

6.4.3.1 Overall Results

We only present the F-statistic results here; the χ^2-statistic showed very similar patterns. The average correlation between the two for each model shows the strong similarity (Table 6.7).

For the F-statistic results, presented in Table 6.8, we report the highest F-statistic in the N-best list ($N = 1, 5, 20, 50$) for each model. For models ErrorCoocc and L1Assoc we report the highest F-statistic for each value of k ($k = 1, 2, 3$). The number of occurrences of the error context with the highest F-statistic is given in parentheses after the F-statistic; the highest value for each N is in bold. For MaxWeightClique-L1, we also note the language of the graph from which the highest score was derived.

We note by comparing Table 6.8 with Table 6.6 that for each error type except for MD, it is possible to find an error context that is more strongly associated with L1s than is the raw error type alone. For MD this is not surprising, as its frequency of occurrence is very strongly linked to the L1, as noted in Table 6.5 and §6.4.2.2.[19] (For the error type MT also, no model produces an error context more strongly associated with the L1 for the single best choice where $N = 1$, but does for larger values of N.)

With respect to the individual models, the simple ErrorCoocc scores highly, giving the best result about half the time, and the best results can occur for any of $k = 1, 2, 3$. The number of instances returned for each error plus error context is larger than for the other models as well, which is not surprising as the model aims to find contexts strongly associated with the errors rather than with L1s. However, these are then likely to be features that are fairly common across L1s; we look at some examples in §6.4.3.2.

L1Assoc performs fairly poorly on our evaluation measure, although in many cases it does find an error context more strongly associated with the L1 than just the raw error type. Counts are also lower. Also, for this model, $k = 2, 3$ are always worse than $k = 1$: bringing in a second context feature reduces the number of occurrences to such an extent that the F-statistic can drop dramatically. This is probably in part an artefact of the size of the FCE corpus (and particularly our FCEsub subcorpus): these features derived from the Toefl11 corpus just do not occur sufficiently often in our evaluation corpus (and in fact there are often large numbers of zero occurrences for $k = 2, 3$).

[19] The fact that determiner errors are very widely studied in terms of analysing cross-linguistic influence suggests a broad consensus that they vary strongly with L1. In addition to Diéz-Bedmar and Papp (2008), a sample of other studies includes Parrish (1987), Young (1996) and Ionin and Montrul (2010).

Error	Context	Example sentences
DJ	JJ, NN1	Basically/RR ,/, I/PPIS1 helped/VVD them/PPHO2 liaise/VV0 with/IW the/AT local/JJ police/NN and/CC get/VV0 some/DD <ns type="DJ">electronical<c>**electronic/JJ**</c></ns> **equipment/NN1** that/CST they/PPHS2 needed/VVD.
		The/AT **show/NN1** will/VM be/VB0 at/II the/AT Central/JJ **Exhibition/NN1** Hall/NP1 and/CC it/PPH1 will/VM be/VB0 <ns type="DJ">opened<c>**open/JJ**</c></ns> until/ICS 7/MC.
UV	TO_VV0_II, NNL1, II, NN2, VV0_II	I/PPIS1 used/VMK **to/TO** <ns type="UV">be</ns> **play/VV0 in/II** the/AT **school/NNL1** team/NN1 . . . and/CC our/APP$ team/NN1 was/VBDZ one/MC1 of/IO the/AT best/JJT basketball/NN1 **teams/NN2** . . .
DN	XX, XX_VV0, VM_XX_VV0, NN1	Never/RR the/AT less/DAR ,/, in/II summer/NNT1 we/PPIS2 **can/VM n't/XX resist/VV0** such/DA <ns type="DN">hot<c>**heat/NN1**</c></ns>!
		. . . I/PPIS1 think/VV0 you/PPY should/VM have/VH0 a/AT1 <ns type="DN">baby-parking<c>kindergarten/NP1</c></ns> ,/, in/II fact/NN1 a/AT1 certain/JJ number/NN1 of/IO women/NN2 **could/VM n't/XX see/VV0** the/AT **Festival/NN1** because/CS of/IO their/APP$ sons/NN2.
MD	VBZ_RG, RG_JJ_NN1	The/AT first/MD and/CC most/RR important/JJ thing/NN1 **is/VBZ that/RG modern/JJ technology/NN1** has/VHZ made/VVN our/APP$ life/NN1 easier/JJR ,/, for/IF instance/NN1 <ns type="MD"><c>the/AT</c></ns> rice/NN1 cooker/NN1 is/VBZ a/AT1 great/JJ invention/NN1 . . .

Figure 6.8: Examples for sample error types and specific error contexts. Error contexts are bolded.

MaxWeightClique also performs fairly poorly. However, in many cases it also finds an error context more strongly associated with L1 than the raw error type alone (DN, MT, UD, UT, UV), even if not always for $N = 1$, and it has intermediate counts of occurrences.

MaxWeightClique-L1 gives the best results in the other half of the cases where ErrorCoocc does not. The error contexts that it finds, however, are very specific, often to a single language (as might be expected by its definition) with very small numbers of counts.

6.4.3.2 Some Examples

We look at some examples in Figure 6.8, to illustrate both interesting error contexts found and areas where the models do a poor job. In these sample sentences, only errors of interest are retained and highlighted.

The DJ error with context { JJ, NN1 } illustrates the top result found under the ErrorCoocc model for $N = 20$. In the first sentence the model seems to find a useful pattern: the adjective that is at the centre of the error occurs in the context of a singular noun. On the other hand, the second sentence illustrates a problem: because the range of the context is the whole sentence, frequent features such as NN1 will occur a lot in other parts of the sentence that have no apparent relation to the actual error. The ErrorCoocc model is thus likely to be picking up false positives by virtue of the relatively high frequencies of its error contexts.

The UV error with context { TO_VV0_II, NNL1, II, NN2, VV0_II } illustrates the top result found under the MaxWeightClique-L1 model for $N = 5$. This is very specific, and its three instances only appear in Turkish. But all three are similar errors from different documents, so it appears likely to be a genuine pattern, although the NN2 seems only to have a tenuous connection.

The DN error with context { XX, XX_VV0, VM_XX_VV0, NN1 } illustrates the top result found under the MaxWeightClique-L1 model for $N = 50$. A number of this reasonably sized set are similar to the first sentence, where the context appears interesting. In this example, *hot* is used for *heat*; the other examples of this type are from Spanish and Italian (similarly, e.g., *live* for *life*), where

error	N	ErrorCoocc	L1Assoc	MaxWeightClique	MaxWeightClique-L1
DJ	1	2.78(274) / **3.19**(227) / 2.95(158)	1.59(31) / 1.59(31) / 0.81(6)	0.99(15)	3.08(2) [ger]
	5	**3.60**(268) / 3.19(227) / 3.02(148)	2.19(12) / 1.59(31) / 0.81(6)	1.74(41)	3.24(2) [chi]
	20	3.72(194) / 3.33(163) / **4.02**(93)	2.53(70) / 1.59(31) / 1.36(1)	2.34(24)	3.50(5) [ita]
	50	3.72(194) / 3.39(114) / **4.02**(93)	2.58(107) / 1.59(31) / 1.59(31)	2.48(18)	3.84(3) [ita]
DN	1	0.77(268) / 1.63(185) / 1.73(119)	1.09(40) / 1.09(40) / 0.70(7)	1.26(63)	**3.24**(2) [chi]
	5	1.80(191) / 2.29(153) / 2.54(142)	1.25(5) / 1.36(1) / 1.36(1)	1.26(63)	**3.24**(2) [chi]
	20	2.34(86) / 2.69(144) / 2.95(113)	2.04(26) / 1.36(1) / 1.36(1)	1.76(30)	**3.24**(2) [chi]
	50	2.86(61) / 3.16(120) / 2.95(113)	3.89(4) / 2.75(2) / 2.75(2)	3.41(18)	**4.27**(10) [spa]
MD	1	**14.28**(1319) / 9.09(985) / 6.38(753)	5.83(198) / 5.83(198) / 0.54(2)	3.07(297)	4.05(91) [kor]
	5	**14.28**(1310) / 12.18(769) / 6.75(582)	8.20(268) / 5.83(198) / 1.93(3)	5.83(198)	5.83(198) [kor]
	20	**14.41**(850) / 12.18(769) / 6.82(593)	8.20(268) / 5.83(198) / 2.60(36)	5.83(198)	5.83(198) [kor]
	50	**14.41**(850) / 12.18(769) / 7.99(483)	8.36(831) / 5.83(198) / 5.83(198)	5.83(198)	6.47(110) [kor]
MT	1	**3.34**(794) / 3.00(666) / 3.02(485)	1.85(79) / 1.85(79) / 1.55(13)	1.70(61)	2.48(20) [chi]
	5	3.34(794) / 3.46(478) / 3.37(378)	2.54(101) / 1.85(79) / 1.55(13)	2.14(64)	**4.47**(3) [chi]
	20	4.44(295) / 3.64(375) / **4.60**(294)	4.44(295) / 3.11(25) / 3.11(25)	2.79(44)	4.47(3) [chi]
	50	4.50(277) / **5.21**(247) / 4.72(215)	4.44(295) / 3.86(33) / 3.11(25)	4.54(74)	4.61(3) [ger]
UD	1	0.69(679) / 1.05(475) / **2.08**(334)	1.45(62) / 1.45(62) / 0.73(10)	0.64(47)	1.54(20) [ger]
	5	1.70(405) / 1.17(452) / 2.08(334)	1.59(26) / 1.45(62) / 1.36(1)	1.45(62)	**3.54**(9) [chi]
	20	2.08(223) / 2.11(360) / 2.32(276)	3.41(51) / 1.45(62) / 1.36(1)	1.90(29)	**3.93**(3) [ita]
	50	3.27(112) / 3.01(188) / 2.33(198)	3.41(51) / 1.54(4) / 1.54(4)	2.85(66)	**4.06**(3) [ita]
UT	1	0.14(548) / 0.45(414) / 1.12(259)	1.01(51) / 1.01(51) / 0.43(1)	0.81(35)	**3.06**(2) [ger]
	5	0.82(368) / 1.16(321) / 1.58(249)	2.28(23) / 1.36(1) / 1.36(1)	1.01(51)	**4.10**(3) [tur]
	20	1.51(351) / 1.77(275) / 1.89(225)	2.91(51) / 1.53(6) / 1.36(1)	2.58(45)	**4.10**(3) [tur]
	50	2.25(112) / 2.66(201) / 3.18(178)	2.91(51) / 1.53(6) / 1.36(1)	2.58(45)	**4.10**(3) [tur]
UV	1	0.88(260) / 0.97(186) / 1.18(119)	1.06(15) / 1.06(15) / 1.29(2)	1.49(28)	**2.53**(2) [jap]
	5	2.22(175) / 2.21(162) / 1.68(109)	2.29(8) / 1.29(2) / 1.29(2)	1.49(28)	**4.09**(3) [tur]
	20	2.25(125) / 2.82(127) / 3.13(96)	3.22(8) / 1.52(1) / 1.52(1)	2.38(15)	**4.09**(3) [tur]
	50	2.56(61) / 3.01(101) / 3.13(96)	3.22(8) / 1.52(1) / 1.52(1)	2.38(15)	**4.63**(3) [chi]

Table 6.8: Results for the chosen error types under the four proposed models. All error types and models report the best F-statistic for the selected error context and frequency within the top N ($N = 1, 5, 20, 50$). ErrorCoocc and L1Assoc give the best score for the set of k features ($k = 1, 2, 3$). MaxWeightClique-L1 also notes the language graph with the best result.

the error seems to be connected to words where the English derivational morphology is not simply affixation. However, there are some like the second sentence, where (as for the DJ error) the error context appears in a different clause, and are likely irrelevant.

The MD error in the last row we examine because (a more complex version of) it was the focus of the case study in Yannakoudakis et al. (2012), which from the examples of that paper looked quite convincing as an error context of relevance to SLA research. However, it and the related examples of Yannakoudakis et al. (2012) were not in the publicly available corpus,[20] and in fact there is only one example of this error and context in the whole FCE corpus, illustrating the issue of data sparsity. Further, this example also illustrates the issue of tagging error: *that* is tagged as RG (degree adverb) where it should be CST.

So as might be anticipated from the frequency numbers in Table 6.8, the MaxWeightClique-L1 model produces context that looks interesting from an SLA perspective, but is relatively limited in scope; the ErrorCoocc model produces a much larger set of candidates, and can successfully find error context such that they behave differently with respect to the L1s according to the ANOVA F-statistic, but produces false positives. Overall, a recurring issue illustrated for all models by the examples is the proposal of error context far away from any likely relevance to SLA.

6.4.4 Concluding Remarks and Future Work

In this section, prompted by work on using computational visualisation techniques to help SLA researchers form hypotheses about errors and the environments in which they are made, we have defined a new task for finding interesting contexts for errors that vary with the native language of the speaker. We proposed four models, ranging from one based on simple error-feature co-occurrence statistics to one based on the maximum weighted clique on an L1-specific feature association graph; these all managed to find contexts that were more strongly associated with L1s than the raw errors alone, and produced (albeit with many false positives in the case of the simple model) some error contexts that look potentially useful for SLA.

6.4.4.1 Future Directions for Error Context Extraction

This research is largely intended to prompt more work on applying NLP techniques to SLA more broadly. As such, there are many ways in which the work could be further developed. First, to get rid of obviously incorrect cases, the size of the area over which the feature-feature and feature-error scores are calculated could be restricted, perhaps to the relevant clause or a certain window size. Second, it may not be the case that the ANOVA F-statistic or χ^2 are the best evaluation measure: in medical work, for example, there is the notion of clinical significance, which takes effect size into account and is often more relevant to the practitioner than statistical significance. Similarly, the current features may not be the most meaningful. As part of this, an important step would be to bring in SLA researchers, to assess proposed error contexts and look at what evaluation measures best relate to this. The role of the present work would then be to rule out models for producing error contexts (like L1Assoc) that produce weaker results in hypothesis testing: it would thus be complementary to the visualisation work from which it stems, guiding SLA researchers away from unproductive areas of the space of possible hypotheses. And third, the size of the corpus is (as always) an issue: as these error-annotated corpora are few and far between, a semi-supervised approach or

[20] We assume that the multiple examples come from the larger CLC corpus.

one that in some way incorporated unannotated data would be useful, perhaps using some of the extensive recent work on error annotation.

Furthermore, there are potentially other methods to increase the accuracy of our analysis. One such approach is to limit our analysis to errors and features that co-occur within the within the same clause, rather than just the sentence. This assumes that errors and features that occur across clause boundaries are not related, and that the accuracy of our analysis can be improved by excluding them. One possible flaw with this method is its dependence on the accurate parsing of clauses within the texts, which is not guaranteed in our learner texts that often contain significant amounts of syntactic and grammatical errors. Given the possible difficulties with accurate parsing of clauses, a simplified version of this approach would be the use of a fixed window size around errors, where only features that are present within this window would be considered as related.

6.5 Chapter Summary

In this chapter we:

- considered how knowledge obtained from NLI models can inform SLA research
- proposed a method for extracting and ranking overused/underused features per L1
- analyzed some of the most discriminative linguistic features extracted using our method
- showed that the features correspond to phenomena and hypotheses from the SLA literature
- defined a new task for finding error contexts that vary by the native language
- developed new graph-theoretic models for identifying potential SLA hypotheses centred on error contexts
- demonstrated that they find contexts that are more strongly associated with L1s than raw errors alone

Chapter 7

Native Language-based Text Segmentation

In the previous chapter we found that features used in the NLI classification task have potential to be used in other tasks, there connected to SLA. In this chapter we look at how classification models learned from NLI can be applied to the task of tracing linguistic influence in L2 texts that contain L1 influence from more than one language. This may be a document with language productions of more than one author or the work of a single multilingual author. We look at one way of framing this type of question as an NLP task, as a variant of text segmentation where the goal is to segment texts based on L1 influence. We position this with respect to (the small amount of) other work on text segmentation that is not based on topic and then explore a range of Bayesian models along the lines of Eisenstein and Barzilay (2008) for addressing our tasks.

Chapter Contents

7.1 Motivation

NLP methods have been used to perform different types of text analysis, such as topic modelling and sentiment analysis. One novel type of textual analysis involves the identification of sub-sections of a text that show influence from different linguistic backgrounds. This could involve a text composed by multiple authors or the work of a single multilingual author.

An example of such a multilingual author is Joseph Conrad, known for having written a number of famous English-language novels: *Almayer's Folly*, *Heart of Darkness*, *Lord Jim* and *Nostromo*, amongst others. Conrad (born Józef Teodor Konrad Korzeniowski; 1857–1924) was born in Poland and moved to England at the age of 21. He was raised speaking Polish and French and in addition to learning English, he also had knowledge of Latin, German and Greek.

His writings have been the subject of much literary analysis, with one particular direction of such research being the identification of likely L1 influences on his English writing, given that he was fluent in both Polish and French.

In *The Linguistic influence of Polish on Joseph Conrad's style*, Morzinski (1994) investigates the influence of the Polish language on Conrad's work, suggesting that it had great impact on his writing style. To this end, she points to different types of influences in his work ranging from some very abstract literary characteristics ("that Slavonic defeatism with which all his writing is permeated" in the words of Coleman (1931) that Morzinski cites) but also more concrete syntactic and morphological characteristics (*e.g.* "*Several had still their staves in their hands*" where the adverb *still* is awkwardly placed, although it follows the correct Polish word order) of English sentences that seemingly invoke Polish-like sentence construction and word order.

Morzinski identifies in particular verb forms (*e.g.* Polish emphasis on aspect rather than tense), voice (Polish third reflexive voice) and placement of adverbs (partly as a consequence of differences in morphology between Polish and English). Several examples extracted from Conrad's writings by Morzinski are listed below.

(7.1) One gets *sometimes* such a flash of insight.
Miewa sie taki blysk intuicji.

(7.2) I knew *once* a Scotch sailmaker [...]
Znalem **raz** szkockiego zaglomistrza [...]

(7.3) like a flash of lightning *came* to her the *reminiscence* of that despised and almost forgotten civilization she had only glanced at in her days of restraint, of sorrow, and of anger [...]

(7.4) But the resources of my sagacity I did not review.

(7.5) In that equipage they hawked *from town to town* the watermelons of central Russian.

In 7.1 we see an example where the aspectual meaning of the verb in Polish precludes the need for an adverb which would be required for an English equivalent. Here Conrad has placed it after the verb, following standard Polish word order. Similarly, standard Polish word order generally places adverbs of frequency directly after the main verb, something often observed in Conrad's English writing as demonstrated in example 7.2.

Due to its rich morphology, Polish is also a free word order language where the most significant or emphasized items are often placed first. Accordingly, inverted English word order is also not

uncommon in Conrad's works. Example 7.3 illustrates a case where the verb and subject are inverted. Similarly, an OSV sentence is shown in example 7.4.

Morzinski also notes that prepositional phrases can appear anywhere in Polish sentences, while they strongly prefer to remain near their heads in English. These result in some very stylistically marked sentences in Conrad's works. The final example, 7.5, shows such an interrupting prepositional phrase in a sentence.

These are but a few examples of the stylistic features detailed by Morzinski. Peters (2013) writes of Morzinski's work:

> She also contends that attempts to maintain nuances of Polish along with its verbal and reflexive features interfere at times with Conrad's stylistic fluency. Morzinski identifies two particular Polish features that arise in Conrad's English: his use of time with verbs and adverbs. She suggests that the less complex grammatical temporal relationships as well as certain semantic features of Polish cause Conrad to employ, for example, temporal verb constructions and adverbs in ways unusual for native English speakers. Similarly, Conrad's extensive use of intransitive verbs imply his efforts to put into English certain features of Polish (such as reflexive passives, formal passives, and impersonals) that exist differently in English. She feels that this factor may have led to what many have seen as Conrad's impressionistic style. Morzinski goes on to argue that the influence of Polish on Conrad's written English seems to have decreased over the course of his career.

Similarly, Hervouet (1990) argues for the influence of French literature and language on Conrad's works in *The French Face of Joseph Conrad*.

Such analyses of Conrad's writings have been performed manually and this raises the question: can NLP and computational models — unsupervised methods, in particular — assist in performing such tasks? This would require models capable of detecting a particular change in authorial style and is related to other types of stylistic change such as plagiarism detection and authorship attribution.

The detection of authorship changes in a text is another area where such text segmentation methods can be helpful. Documents with potentially more than one author are one scenario where identifying text segments with differing writing styles could be potentially useful. This can also be applied to documents which are purportedly composed by a single author but actually contain contributions from multiple authors. One application of such methods has been in Bible authorship attribution, which we discuss in §7.2.2.

Other types of authorial and stylistic change detection tasks include plagiarism detection (§7.2.4) and poetic voice change detection (§7.2.3). These applications can be thought of as being *intrinsic*, where no reference corpus is available and unsupervised methods must be used. In addition to this, the methods must be primarily focused on detecting syntactic changes and shifts as opposed to just topic changes. This means that non-lexical features, like those we have seen in previous chapters, must be applied for this task.

The primary focus of this chapter is to consider how the analysis and identification of different L1 influences in a text can be framed as a computational linguistics task. As with the above-mentioned work, we propose to conceptualize it as a variety of text segmentation where we aim to split a text into regions that display differing L1 influence.

To this end we investigate Bayesian segmentation models, drawing on the previous work defining this, in particular Eisenstein and Barzilay (2008). This is done in conjunction with the generation of artificial documents with influences from multiple L1s.

The remainder of this chapter is organized as follow. In the next section we begin by summarizing some past NLP work that has tackled similar problems. We then outline our experimental setup, followed by details of our experiments and results. We conclude with a discussion and possible directions for future work.

7.2 Related Work

We will first introduce topic segmentation, which is the most widely applied type of segmentation, along with the relevant evaluation measures. This is followed by several other segmentation tasks which are not based on topic: Bible authorship, poetry voice detection and plagiarism detection.

7.2.1 Topic Segmentation

The most widely-researched text segmentation task is that of topic segmentation, where the goal is to divide a text into topically coherent segments. The concept of *lexical cohesion* is an important aspect of research in this area. Lexical cohesion refers to the principle that text is not formed by a random set of words and sentences but rather logically ordered sets of related words that together form a topic. The cohesion refers to the fact that sentences "stick together" to form a single topical unit. In addition to the semantic relation between words, other methods such as back-references and conjunctions also help achieve cohesion. A detailed treatment of cohesion can be found in the work of Halliday and Hasan (1976). Additionally, a more succinct summary of lexical cohesion can also be found in Morris and Hirst (1991, §1).

Lexical cohesion is not just limited to relations between word pairs but occurs among sequences of related words, known as *lexical chains*. These chains provide clues about the cohesion and structure of a text and can be considered a "good indication of the linguistic segmentation" (Morris and Hirst, 1991). For example, TOEFL11 essays written in response to a prompt about understanding concepts and learning facts[1] might contain something like the following two lexical chains: {`education, learning, understanding`} and {`fact, theory, idea, concept, knowledge`}. Morris and Hirst (1991) were the first to propose a method for the computation of such chains, using a thesaurus as the underlying knowledge base. They note:

> When a lexical chain ends, there is a tendency for a linguistic segment to end, as the lexical chains tend to indicate the topicality of segments. If a new lexical chain begins, this is an indication or clue that a new segment has begun. If an old chain is referred to again (a chain return), it is a strong indication that a previous segment is being returned to.

Given its relevance to topic segmentation, a number of different models have been proposed based on the principle lexical cohesion. The task here is to identify points in a text – usually at sentence boundaries – where lexical coherence is low and then predict the topic change boundaries according to

[1]The specific prompt, referred to P7 in the corpus and listed in Table 2.4, is: Do you agree or disagree with the following statement? It is more important for students to understand ideas and concepts than it is for them to learn facts. Use specific reasons and examples to support your answer.

these local minima. This is generally achieved by selecting points that minimize the level of coherence between sequential segments.

7.2.1.1 TextTiling

The *TextTiling* algorithm (Hearst, 1994, 1997) is one early and popular approach to text segmentation. It is an unsupervised and domain-independent method for automatic subdivision of a text into multi-paragraph units that each represent a topic. The algorithm relies on lexical frequency and distribution information to determine topic boundaries, assuming that each topic has its own vocabulary and that large shifts in this vocabulary usage correspond to topic shifts.

Based on this, a tokenized document is first divided into lexical blocks (*i.e.* pseudo-sentences) of a predefined size. Next, the similarity between all adjacent lexical blocks is measured using a cosine similarity metric. Scores are assigned to all block boundaries and the topic boundaries are then set at the minima along this sequence. These boundaries are then adjusted to coincide with the paragraph boundaries in the original document. The algorithm automatically detects the number of segments in each document.

TextTiling has been shown to work well on segmenting lengthy, expository documents. It was evaluated by comparing the algorithm against reader judgements. Seven readers were asked to mark topic change paragraph boundaries in 13 articles. Agreement among the judges was imperfect and only those boundaries marked by at least three readers were retained as "true" boundaries for evaluation. The method was evaluated according to how many true boundaries it selected out of the total (precision) and how many true boundaries out of the total possible were marked (recall). The TextTiling segmentation (66% precision, 61% recall) outperformed a baseline system that placed random boundaries (44% precision, 37% recall) and its segmentations also matched closely with those of human annotators (81% precision, 71% recall).

Although precision and recall are standard measures in information retrieval, researchers soon noted that their use for evaluating segmentation are problematic. One issue is the trade-off between precision and recall: increasing one usually decreases the other. Additionally, they are not sensitive to near-misses: missing a boundary by either a minimal or wide margin results in a score of 0. Beeferman et al. (1997) note that "an algorithm that places a boundary a sentence away from the actual boundary every time actually receives worse precision and recall scores than an algorithm that hypothesizes a boundary at every position." These shortcomings led to the proposal of specific text segmentation metrics, which we discuss in the next section.

7.2.1.2 Evaluation Metrics

There exist two standard and widely used text segmentation metrics that compare predicted segmentation boundaries against a gold standard reference. Both of these measures are being reported in recent work on topic segmentation and we also follow this practice in our own work.We now briefly describe these two metrics.

P_k Proposed by Beeferman et al. (1997, 1999), this metric is aimed at resolving some of the issues associated with using precision and recall, including partial scoring for near-misses. The method uses a sliding window that is calculated as being half the average size of a segment. This window is moved over the text and at each instance it counts if the element at the start of the probe window

is predicted as being in the same segment as the last element and if this is actually false, *i.e.* the proposed segmentation disagrees with the gold standard. These counts are normalized by the number of measurements to produce a score between 0 and 1. Here a lower score means better performance, with 0 being a perfect segmentation. The authors report P_k values of 0.12 for segmenting broadcast news and 0.19 for segmenting *Wall Street Journal* text. However, it has also been noted that some "degenerate" algorithms – such as placing boundaries randomly or at every possible position – can score 0.5 (Pevzner and Hearst, 2002).

WindowDiff (WD) The P_k has been criticized by Pevzner and Hearst (2002) for applying different penalties for false positive and false negatives in addition to failing to penalize false positives at all in some cases. They propose the *WindowDiff* (WD) metric to remedy these shortcomings. Instead of just testing if two sentences are in the same segment, WD requires that the number of segments between the two sentences be the same in both the hypothesized and gold standard segmentations.

7.2.1.3 Unsupervised Bayesian Segmentation

The work of Eisenstein and Barzilay (2008) is a very different approach compared to TextTiling that has become popular. The authors propose a novel Bayesian approach based on a generative model that assumes that each segment has its own language model. Under this assumption the task can be framed as predicting boundaries at points which maximize the probability of a text being generated by a given language model. An open-source implementation of the method, called BayesSeg, is made available by the authors.

We discuss their model at some length here: it is the basis of many later models (which we briefly note at the end of this section) and also the basis for the work in this chapter.

Their method is based on lexical cohesion — expressed in this context as topic segments having compact and consistent lexical distributions — and attempts to implement this within a probabilistic framework by modelling words within each segment as draws from a multinomial language model associated with that segment.

In Equation 1 of their work they define the observation likelihood as,

$$p(\mathbf{X} \mid \mathbf{z}, \Theta) = \prod_{t}^{T} p(\mathbf{x}_t \mid \theta_{z_t}), \tag{7.1}$$

where $\mathbf{X}$ is the set of all T sentences, $\mathbf{z}$ is the vector of segment assignments for each sentence, $\mathbf{x}_t$ is the bag of words drawn from the language model and Θ is the set of all K language models $\Theta_1 \ldots \Theta_K$. K is assumed to be fixed and known, which is standard in segmentation work. The authors also impose an additional constraint, that $\mathbf{z}_t$ must be equal to either $\mathbf{z}_{t-1}$ (the previous sentence's segment) or $\mathbf{z}_{t+1}$ (the next segment), in order to ensure a linear segmentation.

The key concept here is that the language model corresponding to each segment must concentrate its probability mass on a compact set of words, consistent with the principle of lexical cohesion.

This segmentation model has two parameters: the set of language models Θ and the segment assignment indexes $\mathbf{z}$. The authors note that since this task is only concerned with the segment assignments, searching in the space of language models is not desirable. They offer two alternatives to overcome this: (1) taking point estimates of the language models (Eisenstein and Barzilay, 2008,

§3.1) which is considered to be theoretically unsatisfying and (2) marginalizing them out (Eisenstein and Barzilay, 2008, §3.2), which yields better performance.

Equation 7 of Eisenstein and Barzilay (2008), reproduced here, shows how they marginalize over the language models,

$$\begin{aligned} p(\mathbf{X} \mid \mathbf{z}, \theta_0) &= \prod_{j}^{K} \prod_{\{t:z_t=j\}} p(\mathbf{x}_t \mid \theta_0) \cdot \\ &= \prod_{j}^{K} \int d\theta_j \prod_{\{t:z_t=j\}} p(\mathbf{x}_t \mid \theta_j) p(\theta_j \mid \theta_0) \\ &= \prod_{j}^{K} p_{dcm}(\{\mathbf{x}_t : z_t = j\} \mid \theta_0) \end{aligned} \tag{7.2}$$

The definition of p_{dcm} — which refers to the Dirichlet compound multinomial distribution — follows. This DCM distribution expresses the expectation over all of the multinomial language models, when conditioned on the symmetric Dirichlet prior θ_0,

$$p_{dcm}(\{\mathbf{x}_t : z_t = j\} \mid \theta_0) = \frac{\Gamma(W\theta_0)}{\Gamma(N_j + W\theta_o)} \prod_{i}^{W} \frac{\Gamma(n_{j,i} + \theta_0)}{\Gamma(\theta_0)} \tag{7.3}$$

where W is the number of words in the vocabulary and $N_j = \sum_i^W n_{j,i}$, the total number of words in the segment j. The gamma function $\Gamma(n)$ is an extension of the factorial where $\Gamma(n) = (n-1)!$.

The authors then observe that the optimal segmentation maximizes the joint probability

$$p(\mathbf{X}, \mathbf{z} \mid \theta_0) = p(\mathbf{X} \mid \mathbf{z}, \theta_0) p(\mathbf{z})$$

and assume a uniform $p(\mathbf{z})$ over valid segmentations with no probability mass assigned to invalid segmentations. The hyperparameter θ_0 can be chosen, or can be learned via an Expectation-Maximization process.[2]

The authors evaluate their method on two corpora from different domain: the ICSI corpus of meeting transcripts which is used in speech segmentation, and a new corpus of medical text created by the authors. The meeting corpus has transcripts of 75 multi-party meetings, 25 of which are annotated with segment boundaries. The text corpus is constructed from the contents of a medical textbook[3] composed of 227 chapters where each chapter is segmented into sections with a total of 1136 sections.[4]

Their results show that their method substantially outperforms other segmentation approaches for both the meeting transcripts (their method giving $P_k = 0.34$) and medical textbook datasets ($P_k = 0.26$).

This work presents a significant contribution and a lot of subsequent research has drawn on this methodology for varying purposes.[5] Eisenstein (2009) later expanded this model for hierarchical text segmentation, treating lexical cohesion as a multi-scale phenomenon. The approach was adapted so that each token is modelled as a draw from a pyramid of latent topic models with the structure of

[2] The implementation that the authors provide can do this.

[3] The book, *Clinical Methods: The History, Physical, and Laboratory Examinations*, is available for free download at `http://onlinebooks.library.upenn.edu`

[4] There are an average of 5 segments per chapter, and an average of 140 sentences per chapter.

[5] The paper currently has 125 citations in Google Scholar.

the pyramid being constrained to induce a hierarchical segmentation. Many other studies in the area use the BayesSeg system as a competitive baseline for evaluating their own work. Examples include the Affinity Propagation-based segmentation method of Kazantseva and Szpakowicz (2011) and the content modelling work of Chen et al. (2009).

Others have also extended the Bayesian approach in some way. Jeong and Titov (2010), for example, propose a model for joint discourse segmentation and alignment for documents with parallel structures, such as a text with commentaries or presenting alternative views on the same topic. This is motivated partly by the goal that "revealing relations between the parts by jointly segmenting and predicting links between the segments, would help to visualize such documents" as well as for generating summaries of the aligned segments. They extend the model of Eisenstein and Barzilay (2008) to segment and align parts of a parallel document.

Another example of an extension to the Bayesian model is the work of Du et al. (2013) who modify the Eisenstein and Barzilay (2008) model "into a structured topic model that can capture a simple hierarchical topic structure latent in documents". This is not done to perform hierarchical segmentation, but instead the information from the hierarchies is used to improve the linear segmentation. They also use BayesSeg as a baseline for evaluating their work.

7.2.2 Bible Authorship

Koppel et al. (2011a) consider the task of decomposing a document into its authorial components based on their stylistic properties and propose an unsupervised method for doing so.

Their aim here is to be able to delineate the contributions of each individual author in a multi-author text, where the contributions are entire chapters. The method is unsupervised in the sense that no prior information or other writing samples by the authors are available.

The authors use biblical books as their data, specifically the books of Jeremiah (52 chapters) and Ezekiel (48 chapters) in Hebrew, both of which are generally believed to be single-authored. The chapters in these books are unlabelled. A single artificial corpus was constructed by interleaving chapters of the books of Jeremiah and Ezekiel and the task is to identify those by the same author.

Their first method involves representing each chapter as a bag-of-words (using all words that appear at least $k = 2$ times in the corpus). They then compute the cosine similarity of every pair of chapters in the corpus and apply a clustering algorithm to group the chapters into two clusters. This method does not perform well, with only half the chapters being clustered correctly.

Lexical choice is an important part of authorial style; while one author might prefer using the words *begin* and *end*, another might prefer *start* and *finish*. With this in mind, the authors then propose an extended method that exploits synonym choice. They give as an example the case that in Hebrew there are seven synonyms for the word *fear*, and that different authors may choose consistently from among them. The basic idea here is to identify synonyms based on the hypothesis that authors differ in the proportion of words they use from a set of synonyms (synset).

The authors construct their own synsets using available biblical resources and annotations. They then represent texts by vectors of synonyms[6] and apply a modified cosine similarity measure to compare and cluster these vectors. This results in substantial improvements in the clustering results, demonstrating the utility of synonyms for this task.

[6]The value of a cell is $1/j$ for use of j synonyms in a particular synset.

A key difference between this approach and our task is the assumption that different texts are written by different authors. Our task does not make such an assumption, leaving open the possibility that an author may exhibit different styles of writing as influenced by their different L1s. The method also relies on the extensive historical Biblical scholarship that has identified important synsets for differentiating authors. Relating this to our proposed L1-based segmentation task, this is not the type of differentiating factor that would be normally applied.

7.2.3 Poetry Voice Detection

Others have also attempted to apply unsupervised segmentation methods for literary analysis, specifically the stylistic segmentation of poetry. This was the focus of the work by Brooke et al. (2012) where they perform stylistic segmentation of a well-known poem, *The Waste Land* by T.S. Eliot.

This poem is renowned for the great number of voices that appear throughout the text and has been the subject of much literary analysis (Bedient and Eliot, 1986; Cooper, 1987). These distinct voices, conceived of as representing different characters, have differing tones, lexis and grammatical styles (*e.g.* reflecting the level of formality). The transitions between the voices are not explicitly marked in the poem and the task here is to predict the breaks where these voice changes occur.

The authors argue that the use of generative models is not feasible for this task, noting:

> Generative models, which use a bag-of-words assumption, have a very different problem: in their standard form, they can capture only lexical cohesion, which is not the (primary) focus of stylistic analysis.

They instead present a method based on deriving a curve that captures stylistic change. The local maxima in this change curve represent potential breaks in the text. This stylistic change curve is related to the TextTiling work of Hearst (1994, 1997) which we described in §7.2.1.1, but it has been generalized to apply to any number of features derived from a span of text.[7] The authors propose a number of local features that are internal to the poem (*e.g.* word length, syllable count, part-of-speech tag) as well as extrinsic features that are external to the poem. Some examples of extrinsic features include using the average unigram counts in the 1T Corpus[8] or sentiment polarity from a lexicon.

Initial tests on artificial mixed-style poems, generated by combining stylistically unique poems from distinct authors, show that combining all features provide the best results that are well-above baseline and approximately halfway close to perfect segmentation ($P_k = 0.25$). An interesting finding reported here is the high performance and utility of extrinsic features for this task.

Results for segmenting *The Waste Land* are also above baseline, but lower than those for the artificial poems as this is a much more difficult task which can be challenging even for human readers.[9]

Brooke et al. (2013) extend their investigation of this task by considering a clustering approach. Here they assume an existing initial segmentation of the work and attempt to cluster the segments which correspond to the same voice. They employ the same feature set and apply a slightly modified version of the k-means clustering algorithm. Clusters are compared against a gold standard annotation using the BCubed metric. Results on artificial poems show that the method is effective when

[7] TextTiling only uses a document's vocabulary and performs quantitative lexical analyses to predict the segmentation of the text.

[8] The authors use this as a metric of how commonly a word is used in general.

[9] No accepted gold standard for the segmentation exists. Brooke et al. (2012) enlisted a class of 140 undergraduate students in an English literature course to each segment the poem and used the majority judgement as one component of their gold standard.

stylistic differences are clear and pronounced. For *The Waste Land*, however, the results are far lower and the method less effective. This highlights the limitations of such methods and they conclude that it is likely because the differences in the poem are much more subtle:

> Literature, by its very nature, involves combining existing means of expression in surprising new ways, resisting supervised analysis methods that depend on assumptions of conformity. Our unsupervised approach to distinguishing voices in poetry offers this necessary flexibility, and indeed seems to work reasonably well in cases when the stylistic differences are clear. *The Waste Land*, however, is a very subtle text, and our results suggest that we are a long way from something that would be a considered a possible human interpretation.

7.2.4 Plagiarism Detection

Plagiarism detection is the task of analyzing a text to detect portions which are not original. The goal of *external* plagiarism detection is to find similar documents or segments in a text, given a reference corpus of potentially original documents. On the other hand, *intrinsic* plagiarism detection is a more challenging task that does not rely on any additional documents, instead it analyzes a single document with respect to variations and inconsistencies in writing style (Zu Eissen and Stein, 2006).

Plagiarism has been a particular issue for EFL students. Pecorari (2003) investigated potential plagiarism in academic second-language writing, focusing specifically on *patchwriting*, the heavy use of text from a different source with some modification and insertion of additional words and sentences to form a new text. This is considered a form of plagiarism, even in cases where a citation is provided as quotation marks are not used. Pecorari (2003) studied the writings of 17 postgraduate students, comparing their texts against the original sources. Results showed that the texts exhibited features that could be classified as plagiarism, but further analysis and interviews with the authors suggested that there was no deliberate intent to plagiarize.

Another example is the work of Keck (2006) who compared the use of textual borrowing by both native and non-native writers. The classified paraphrases into four categories: Near Copy, Minimal Revision, Moderate Revision, and Substantial Revision. Results showed that L2 writers had significantly more Near Copy paraphrases than their L1 counterparts. Other researchers have also investigated plagiarism with regards to writers from certain countries (Shi, 2006; Abdul-Ameer and Hussein, 2015).

It should also be noted that this research is generally not conducted by people working in NLP, although there is some similarity in techniques. Change curves like the one described in the previous section have also been explored for intrinsic plagiarism detection. Stamatatos (2009) proposes using a style change function, but based only on character n-grams. The method first involves the calculation of a character n-gram *profile* for the whole document by creating a vector of all n-gram frequencies normalized by the text length, with $n = 3$ in this study. This is followed by passing a sliding window of length l over the entire text and comparing the profile of the content within the window against the document's profile using a dissimilarity function.

This enables the creation of a stylistic change curve along the sliding window positions. The authors then use the standard deviation of these values to detect local maxima which most likely correspond to plagiarised passages.

Using the corpus from First International Competition on Plagiarism Detection, the method achieves an F-score of 0.29 in detecting plagiarised passage. This is a promising result considering that this is a much more difficult task compared to *external* plagiarism detection where the task is to find similar documents or segments in a reference corpus. To our knowledge these task have not yet been tackled using a Bayesian approach.

7.2.5 Summary of Related Work

To summarize this related work, we can see that topic segmentation is a widely-researched task where Bayesian methods have proven to be popular and successful. These methods are based on the concept of lexical cohesion and require that distributions in the model exhibit compactness.

On the other hand, segmentation methods have rarely been applied for stylistic segmentation, with no attempts to apply a Bayesian approach for this task: a goal of this chapter is, contra Brooke et al. (2012), to define a generative model for stylistic segmentation.

7.3 Experimental Setup

The main aim of this work is to propose an unsupervised Bayesian approach to text segmentation based on the native language of the writer(s), the kind of stylistic analysis that is the focus of this thesis. We evaluate our Bayesian models on artificial texts generated from learner essays, similar to the previously described work on Bible authorship and poetry segmentation. These generated documents will then be segmented using an unsupervised approach based on generative models, as described in §7.4.

7.3.1 Document Generation and Data

As the task is to segment texts by the author's L1, we want to ensure that we are not segmenting by topic and thus need use texts written by authors from different L1 backgrounds on the very same topic. The TOEFL11 corpus, which contains both L1 and essay prompt metadata, is a suitable dataset for this task.

Next we choose a pair of languages from which text segments will be drawn. For this work the two L1s we selected were German and Italian. The principle behind our choice is that for the first attempt to tackle this task we should choose L1s that we know have the potential to be distinguished based on the NLI classification task discussed in Chapter 3. German is the class with the highest NLI accuracy in the TOEFL11 corpus and Italian also performs very well. Additionally, there is also very little confusion between the two; a binary NLI classifier trained on the language pair achieved 97% accuracy.

7.3.1.1 Document Generation

We then randomly draw documents from this subset of TOEFL11 to form larger composite documents, each of which contains s segments. TOEFL11 essays — each of which is a segment — are randomly drawn while alternating the L1 class after each segment. This is repeated until the maximum num-

ber of segments per document, s, is reached. In this work we experiment with datasets generated with $s = 5$.[10]

7.3.1.2 Data

We generate four distinct datasets for our experiments using the above methodology data from the TOEFL11-TRAIN and TOEFL11-DEV sets of the TOEFL11 corpus. The documents in these datasets, as described below, differ in the parameters used to select the essays for each segment and what type of tokens are used. Tokens can be represented in their original form where the surface tokens are used for performing segmentation. Alternatively, each token may be replaced by its POS tag and this could be later used to extract POS n-grams. The POS representation is motivated by the fact that our work from the preceding chapters evidenced that POS-based features are very useful for capturing L1-based stylistic differences.[11] Our method for encoding n-grams is described below in §7.3.1.3.

TOPICSEG-TOKENS This data is generated by keeping the L1 class constant and alternating segments between two topics. We chose Italian for the L1 class and essays from the prompts "P7" and "P8" are used.[12] This dataset contains a total of 53 artificial documents and is useful for testing segmentation by topic. It will be used to verify that topic segmentation as discussed in Eisenstein and Barzilay (2008) functions as expected for data from this domain: that is, that topic change is detectable.

TOPICSEG-PTB As POS tags are useful features for capturing L1-based sytlistic differences, they can be useful for this task. An insight here is that we can represent the documents at a level other than lexical (Wong et al., 2012). Here the tokens in each text are replaced with their POS tags and the segmentation is performed over this data. In these experiments the tags are obtained via the Stanford Tagger and use the Penn Treebank (PTB) tagset, which we briefly described in §5.8. The same L1 and topics as TOPICSEG-TOKENS are used for a total of 53 documents. Figure 7.1 shows an example of a document with three segments where the PTB POS tags for each token are used. This dataset will be used to investigate, *inter alia*, whether segmentation by topic is also possible on the basis of stylistic features. We would expect not.

L1SEG-PTB This dataset is used for segmentation based on native language, also using the PTB POS tags. In addition to our language pair we also choose a specific topic and then retrieve all documents from the corpus that match these criteria. For this work we chose prompt "P7" since it had the largest number of essays for our chosen L1 pair. This resulted in 57 documents. To generate a document, essays are randomly drawn while alternating the L1 class after each segment. We would expect that this dataset should not be segmentable by topic, as all the segments are on the same topic; the segments should however, differ in stylistic characteristics related to the L1.

L1SEG-CLAWS2 This dataset is generated using the same methodology as L1SEG-PTB, with the exception that the essays are tagged using the RASP tagger which uses the more fine-grained

[10] However, we have also successfully replicated all of our results using $s = 7, 9$.
[11] See previous results in §3.1.3.5 and §5.8.
[12] The descriptions of the prompts were listed in Table 2.4.

```
==========
IN PRP$ NN , WP VBD IN DT NN VBZ RB RB JJ .
PRP VBP IN DT NN IN IN DT NN , NNS CC NN VBP RBR JJ IN TO VB NNS , VBZ IN DT JJ NN DT NN VBZ .
IN PRP$ NN IN DT NN , PRP VBZ RB RBR JJ TO VB NNS CC NNS IN DT NN .
DT JJS NN FW MD VB TO VB NN TO PRP$ NN , VBZ PRP$ JJ NN IN DT NN NN .
PRP VBP VBN NN CC IN JJS IN DT NNS FW VBD , TO VB NNS VBD DT RBS JJ NN IN DT VBG .
IN DT PRP MD VB VBN RB JJ TO VB DT NN .
IN PRP VBP IN TO VB TO DT NN JJ NN , IN PRP VBP RB VBN IN .
PRP MD VB DT NN IN DT NN MD VB IN DT NN NN VBP RB VBN CC FW MD VB DT NNS .
CC IN PRP VBP RB VB DT NN CC NNS IN NN POS NNP PRP MD RB VBN JJ TO VB PRP RB .
PRP VBP VBG IN NN TO DT NNS JJ TO NN NN .
DT NNS VBP RB JJ IN NNS WDT VBP DT NN IN PRP$ NN NN .
CC RB IN PRP PRP VBP JJ VB JJ TO VB CC TO VB WP VBD IN DT JJ NN CC NN .
TO RBR VB WP VBD JJ TO DT JJ JJ NN .
==========
IN PRP$ NN NN PRP VBD DT NN IN IN NNS RB JJ NN VBZ JJ CC RB RB DT NN IN DT JJ NN CC NNS .
IN NN MD VB IN DT NN DT NNS IN VBG VBP VBN PRP$ NNS .
IN DT JJR JJ NN IN PRP MD VB DT NN IN DT JJ NNS , DT NN VBZ DT JJ NN IN DT NN IN DT NN .
RB PRP CC PRP VBZ DT NN IN DT NN IN DT NN CC RB DT JJ JJ NNS .
DT NN VBZ RB MD VB JJ NNS , WDT RB RB VBZ RB VB IN DT NN IN DT JJ NN .
IN NN IN VBG IN DT NN IN DT NN PRP VBP RB VB TO VB JJ JJ NN .
NN DT JJ NN IN JJ NNS CC JJ NN TO DT NN VBZ DT NN IN DT NN IN PRP CC PRP VBZ TO VB DT NN .
RB IN NN IN NN NN CC RB IN PRP VBZ VBN IN DT NN TO VB VBN IN NN .
IN NN DT JJ NN IN DT JJ NN IN NN NN NN , NNP , IN JJ NN IN JJ NN .
CC RB IN DT NN IN NN PRP IN NN PRP VBZ IN DT JJ NN CC JJ NNS WDT NN IN NN VBZ JJ .
IN PRP PRP$ RBR JJ TO VB NNS IN DT JJ PRP VBP JJ IN CC WP VBZ JJ IN PRP$ VBG NN RB IN .
==========
EX VBP RB VBN VBG NNS IN DT NN IN VBG NNS CC IN NN DT NN IN NNS IN VBG .
IN PRP$ NN DT NN DT NN MD VB PRP$ VBG VBZ RB RB JJ IN DT JJ NN IN NN .
IN NN VBG NN VBZ DT NN IN JJ JJ NNS .
DT VBP NNS NN VBP TO VB VBN , PRP MD RB VB VBN IN DT NN IN DT NN .
RB EX VBZ DT NN IN DT NNS TO VB PRP .
IN DT JJ NN VBG NNS , DT VBP DT NN IN NNS CC NNS DT NNS VBP TO VB IN NN TO VB JJ .
DT MD DT VB VBN IN
CC MD DT NN IN NN NN NN IN PRP$ VBG RB IN VBG NNS .
PRP RB VBZ TO VB IN NN DT NNS IN JJ NNS .
CC MD DT NN IN NNS VB DT JJ NN RB IN VBG JJ IN DT JJ NNS CC NNS .
EX VBZ RB DT NN TO VB DT NN IN JJ NNS CC NNS WDT RB VBP RB VB VBN .
DT JJ NN VBZ RB DT NN IN NN CD NN VBZ .
DT NNS MD VB NN PRP VBP VBN IN PRP$ NNS RB , CC PRP$ JJ IN PRP TO VB NNS IN .
NNS VBP NNS IN VBG NNS , CC VBP JJ TO VB JJ NNS IN NNS PRP VBP VBN RB .
IN PRP MD RB VB VBN IN RB VBG NNS , CC VBG NNS VBZ DT NN TO JJ VBG .
DT NN MD VB DT JJ NN IN DT CD NNS .
VBG IN WP DT NN IN VBG VBZ CC RB VBG IN WP PRP POS NNS CC NN VBP .
```

Figure 7.1: An example of a document with three segments where the PTB POS tags for each token are used.

CLAWS2 tagset, as we described earlier in §5.8. This also resulted in 57 documents. The CLAWS2 tagset performed better in the NLI classification task and we will use this dataset to investigate if there are any differences in this task.

7.3.1.3 Encoding n-gram information

In this section we describe our method for incorporating n-gram information into our topic models. We do this as our preceding work, *e.g.* §5.8, has shown that n-grams outperform unigrams for NLI.

For lexical items, Lau et al. (2013) investigated the importance of n-grams within topic models. They note that in topic modelling each token receives a topic label and that the words in a collocation — *e.g. stock market*, *White House* or *health care* — may receive different topic assignments despite forming a single semantic unit.

They then postulate that encoding such bigrams as single tokens guarantees that these units receive a single label, potentially helping enhance the performance of topic models. This is the first work to assess the impact of n-gram tokenization for LDA topic models, and the authors propose the pre-extraction and tokenization of these bigram collocations by first identifying the top collocations using a t-test. Through extensive studies using four corpora, they experiment with tokenizing the top 1k, 10k and 100k collocations, reporting that using up to 1,000 of the top collocations provides the best improvement over the unigram bag of words approach.

Given the parallels between our segmentation task and this topic modelling work, we also apply the tokenization approach to our data in order to enable the L1Seg model to encode sequences of POS tags. We do this by implementing a preprocessing step that converts each sentence within each document to a set of bigrams or trigrams, where each n-gram is represented by a single token.[13] This procedure greatly increases the number of tokens per sentence and we posit that by explicitly encoding POS tag sequences as individual tokens the segmentation model will be able to better distinguish stylistic changes in the same way that POS n-grams (with $n = 2, 3$) improve NLI accuracy.

7.4 Segmentation Models

For segmentation we adopt the unsupervised Bayesian method of Eisenstein and Barzilay (2008), as described in §7.2.1.3. An open-source implementation of the method, called BayesSeg, is made available by the authors.[14] Datasets generated as part of our experiments are then segmented and we report results according to the metrics described in §7.2.1.2. We experiment with four models for segmentation, described below.

7.4.1 TopicSeg

Our first model is exactly the one proposed by Eisenstein and Barzilay (2008). The aim here is to look at how we perform at segmenting learner essays by topic in order to confirm that topic segmentation works for this domain and these types of topics. This is something that has been shown to be quite feasible by previous research (Beeferman et al., 1999; Eisenstein and Barzilay, 2008) and we aim to evaluate how the different document types perform. We apply this model to the TopicSeg-Tokens

[13] *e.g.* the trigram `DT JJ NN` becomes a single token: `DT-JJ-NN`

[14] `http://groups.csail.mit.edu/rbg/code/bayesseg/`

and TopicSeg-PTB datasets where the texts have the same L1 and boundaries are placed between essays of differing topics.

7.4.2 L1Seg

Our second model is based on that of Eisenstein and Barzilay (2008), but modified slightly with a revised generative story. Where they assume a standard generative model over words with constraints on topic change between sentences, we make make minor modifications to adapt the model for our task. The standard generative story (Blei, 2012) — an account of how a model generates the observed data — usually generates words in a two-stage process:

1. For each document, randomly choose a distribution of topics
2. For each word in the document:
   (a) Assign a topic from those chosen in step 1
   (b) Randomly choose a word from that topic's vocabulary

Such generative models, although initially developed for language-related tasks such as topic modelling, have been adapted to find trends in images, genetic data and social networks (Blei, 2012). Here we modify this story to be over part-of-speech data instead of lexical items. By using this syntactic data we aim to segment our texts based on the native language of the author for each segment, with each "topic" now representing an L1. For this model we only make use of the L1Seg-PTB dataset since POS tags have been shown to capture inter-L1 differences in syntax. This model also applies to POS n-grams, as we described in §7.3.1.3.

7.4.3 L1Seg-Compact

It is also not obvious that the same methods that produce compact distributions in standard lexical chains would also work for POS data, particularly if extended to POS n-grams which can result in a very large number of tokens. In this regard Eisenstein and Barzilay (2008) note:

> To obtain a high likelihood, the language models associated with each segment should concentrate their probability mass on a compact subset of words. Language models that spread their probability mass over a broad set of words will induce a lower likelihood. This is consistent with the principle of lexical cohesion.

Eisenstein and Barzilay (2008) discuss this within the context of topic segmentation. For example, a topic segment related to the previously mentioned essay prompt P7 might concentrate its probability mass on the following set of words: {`education, learning, understanding, fact, theory, idea, concept, knowledge`}. However, it is unclear if this would also would happen for POS tags; there is no syntactic analogue for the sort of lexical chains important in topic segmentation. It may then turn out that using all the POS tags as we did in the previous model would not achieve a strong performance. Therefore additional approaches for compacting the set of tags used by the generative model, using knowledge from the classification task, could potentially be necessary for enhanced performance.

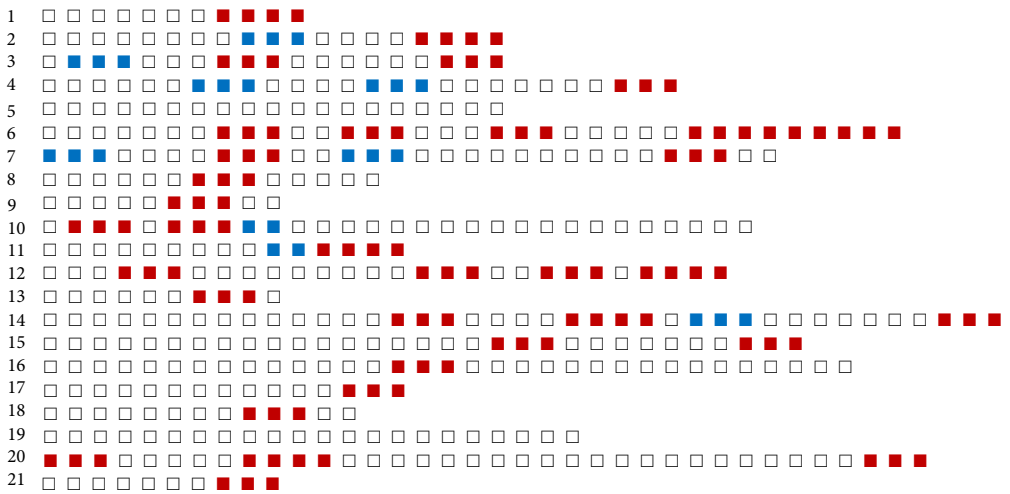

Figure 7.2: A visualization of sentences from a single segment. Each row represents a sentence and each token is represented by a square. Token trigrams considered discriminative for either of our two L1 classes are shown in blue or red, with the rest being considered non-discriminative.

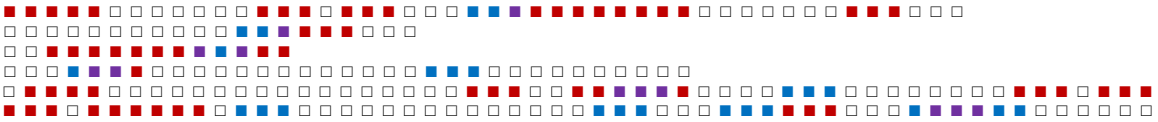

Figure 7.3: Tokens that are part of discriminative trigrams for both L1 classes are shown in purple here.

Discarding Non-Discriminative Features One approach that could possibly overcome these limitations is the removal of non-discriminative features from the input space. This would allow us to encode POS sequence information via n-grams while also keeping the model's vocabulary sufficiently small. Doing this requires the use of extrinsic information for filtering the n-grams. The use of such extrinsic information has proven to be useful for other similar tasks such as the poetry style change segmentation work of Brooke et al. (2012), as described in §7.2.3.

We perform this filtering using the discriminative feature lists extracted using the method we proposed in Chapter 6. We extract the top 300 most discriminative POS n-gram features for each L1, resulting in two lists of 600 POS bigrams and trigrams. This list also enables us to visualize the distribution of the discriminative features within our documents. Figure 7.2 shows one such visualization of a single segment where each row represents a sentence and each token is represented by a square. Tokens that are part of a trigram which is considered discriminative for either of our two L1 classes are shown in blue or red.

Note that discriminative trigrams can overlap with each other within the same class (*e.g.* on lines 1 and 2 where two overlapping trigrams form a group of four consecutive tokens) and also between two classes (*e.g.* on lines 10 and 11). This can occur in cases where the last tag in a discriminative trigram overlaps with the first tag in another such n-gram. This is also shown in Figure 7.3 where overlapping elements are displayed in purple.

Cases of overlap across two classes can be potentially problematic since we need to select the most discriminative feature. We resolve such conflicts by using the weights[15] assigned to the features: in cases of overlap, the feature with the higher weight will be chosen.

[15]These weights come from the discriminative model used to extract informative features, see §6.3 for more details.

7.4.4 L1Seg-AsymP

Looking at the distribution of discriminative features in our documents, as in Figure 7.2, one idea is that incorporating knowledge about which features are associated with which L1 could potentially help improve the results.

One approach to do this might be the use of asymmetric priors. We note that features associated with an L1 often dominate in a segment, *e.g.* in Figure 7.2 we observe a preponderance of red squares. Accordingly, priors can represent evidence external to the data that some some aspect should be weighted more strongly: for us, this is evidence from the NLI classification task.

The segmentation models discussed so far only make use of a symmetric prior[16] but later work mentions that it would be possible to modify this to use an asymmetric prior (Eisenstein, 2009), although this does not appear to have been implemented.

Given that they are effective for incorporating external information, recent work has highlighted the importance of optimizing over such priors, and in particular, the use of asymmetric priors. One example is the work of Wallach et al. (2009) on LDA where they report that "an asymmetric Dirichlet prior over the document-topic distributions has substantial advantages over a symmetric prior", with the priors being approximated through hyperparameter optimization. Such methods have since been applied in other tasks such as sentiment analysis (Lin and He, 2009; Lin et al., 2012) to achieve substantial improvements. For sentiment analysis, Lin and He (2009) incorporate external information from a subjectivity lexicon. As is the case here, the sentiment elements draw on external info, there sources like the MPQA subjectivity lexicon. In applying LDA, instead of using a uniform Dirichlet prior for the document–sentiment distribution, they experiment with an asymmetric prior. This value was determined empirically, with the optimal values being 0.01 for positive sentiment elements and 5.0 for negative sentiment ones. This result illustrates that there can be a very large difference between the priors, with one being very strong. Such strong priors can give more weight to other beliefs (for our task, the external NLI information) with respect to the data.

Conceptually, or our task this could at its simplest entail having two asymmetric priors, one corresponding to L1_a and the other to L1_b.[17] Given this, we can assume that segments will alternate between L1_a and L1_b. And instead of a single θ_0, we have two asymmetric priors that we call θ_a, θ_b corresponding to L1_a and L1_b respectively. This will require reworking the definition of p_{dcm} in equation 7.3 in §7.2.1.3.

The next issue is how to construct the θ_a and θ_b. The simplest scenario would require a single constant value for all elements in one L1 and another for all elements in the other L1. Specifically, referring to each element i of the prior with the notation $\theta_a[i]$ or $\theta_b[i]$ as appropriate, and using $\text{discrim}(\text{L1}_x)$ to denote "the ranked list of discriminative n-grams for L1_x", we define

$$\theta_a[i] = \begin{cases} c_1 & \text{if } \theta_a[i] \in \text{discrim}(\text{L1}_a) \\ c_2 & \text{if } \theta_a[i] \in \text{discrim}(\text{L1}_b) \end{cases}$$

and

[16] They use the default BayesSeg prior value which was optimized for topic segmentation by Eisenstein and Barzilay (2008).

[17] A more sophisticated approach would be to have more numerous fine-grained asymmetric priors for each element of L1_a and L1_b: that would correspond in our case to individual POS n-grams, and the asymmetric priors could be set by reference to the weight of those n-grams from the NLI classification model. Given the relatively small size of our dataset and the large number of parameters this would require, we do not investigate this here.

$$\theta_b[i] = \begin{cases} c_2 & \text{if } \theta_b[i] \in \text{discrim}(\text{L1}_a) \\ c_1 & \text{if } \theta_b[i] \in \text{discrim}(\text{L1}_b) \end{cases}$$

We would expect that $c_1 > c_2$ (*i.e.* the prior is stronger for those elements that come from the appropriate ranked list of discriminative features), but these values will be learned.

Adapting Equation 7 of Eisenstein and Barzilay (2008):

$$p(\mathbf{X} \mid \mathbf{z}, \theta_a, \theta_b) = \prod_{\{j_o : j_o \bmod 2=1, j \leq K\}} \prod_{\{t : z_t = j_o\}} p(\mathbf{x}_t \mid \theta_a) \tag{7.4}$$

$$\prod_{\{j_e : j_e \bmod 2=0, j \leq K\}} \prod_{\{t : z_t = j_e\}} p(\mathbf{x}_t \mid \theta_b) \tag{7.5}$$

$$= \prod_{j_o} p_{dcm}(\{\mathbf{x}_t : z_t = j_o\} \mid \theta_a) \prod_{j_e} p_{dcm}(\{\mathbf{x}_t : z_t = j_e\} \mid \theta_b) \tag{7.6}$$

Note that j_o is an index over odd segments and j_e over even ones. K is as usual the total number of segments.

$$p_{dcm}(\{\mathbf{x}_t : z_t = j_o\} \mid \theta_a) = \frac{\Gamma(\sum_k^W \theta_a[k])}{\Gamma(N_{j_o} + \sum_k^W \theta_a[k])} \prod_i^W \frac{\Gamma(n_{j,i} + \theta_a[i])}{\Gamma(\theta_a[i])} \tag{7.7}$$

W is now more generally the number of items in our vocabulary (whether words or POS n-grams). A notational addition here is $\theta_a[k]$ which refers the L1_a prior for the kth word or POS n-gram. There is an analogous p_{dcm} for θ_b.

In principle we would then calculate $p(\mathbf{X} \mid \mathbf{z}, \theta_a, \theta_b)$ twice: once where we assign θ_a to segment 1, and the second time where we assign θ_b. We'd then compare the two $p(\mathbf{X} \mid \mathbf{z}, \theta_a, \theta_b)$, and see which one fits better. In this work, however, we will fix the initial L1: segment 1 corresponds to L1_a and consequently has prior θ_a.[18]

7.4.4.1 Asymmetric Prior Example

In this section we present a brief worked example of how the asymmetric prior would work. As described in §7.2.1.3, there are two ways to create the language models: taking point estimates of the language model or alternatively, marginalizing the language model. Although we work with the marginalized model in our work, we will use the point estimate for this example as they are more intuitive.

Supposing we have have a vocabulary of five words, which are POS tags[19] in our case: { *DT*, *EX*, *IN*, *JJ*, *NN* }; $W = 5$. In our example, let us say that *EX* is associated with L1_a and the remaining elements with L1_b. Starting first with the symmetric case, let $\theta_o = (1, 1, 1, 1, 1)$.

The calculations of the point estimates for the first segment are then:

[18] This requires an extension of the BayesSeg software to support asymmetric priors. We will make this extended version of the code available under the same conditions as BayesSeg.

[19] These could refer to, for example, { *the*, *there*, *of*, *big*, *cat* }

$$\hat{\theta}_{1,DT} = \frac{n_{1,DT}+1}{\sum_{i=1}^{5} n_{1,i}+5}$$

$$\hat{\theta}_{1,EX} = \frac{n_{1,EX}+1}{\sum_{i=1}^{5} n_{1,i}+5}$$

$$\ldots$$

Recall that θ_j is the language model for segment j. Here we use $\hat{\theta}_{j,w}$ to denote the point estimate value for word w in θ_j. The 1 in the first numerator comes from $\theta_0[1]$, in the second numerator from $\theta_0[2]$, and so on.

If we had one single asymmetric prior, say $\theta'_0 = (1, 2, 1, 1, 1)$, then

$$\hat{\theta}_{1,DT} = \frac{n_{1,DT}+1}{\sum_{i=1}^{5} n_{1,i}+\mathbf{6}}$$

$$\hat{\theta}_{1,EX} = \frac{n_{1,EX}+\mathbf{2}}{\sum_{i=1}^{5} n_{1,i}+\mathbf{6}}$$

$$\ldots$$

Extending this to two priors, assume that the prior θ'_0 defined above is our θ_a. Let $\theta_b = (2, 1, 2, 2, 2)$. Then

$$\hat{\theta}_{2,DT} = \frac{n_{2,DT}+\mathbf{2}}{\sum_{i=1}^{5} n_{1,i}+\mathbf{8}}$$

$$\hat{\theta}_{2,EX} = \frac{n_{2,EX}+\mathbf{1}}{\sum_{i=1}^{5} n_{1,i}+\mathbf{8}}$$

$$\ldots$$

In this example $c_1 = 2$ and $c_2 = 1$. It is apparent that under θ_a and consequently for our point estimate $\hat{\theta}_{1,EX}$, the element EX will be favoured, and under θ_b, the other elements. These priors effect a kind of smoothing: the stronger the prior, the less the data has a say.

7.5 Results

In this section we outline the four experiments we conducted using the above-described models and data. The results for all of our experiments can be found in Table 7.1.

7.5.1 Segmenting by Topic

We begin by testing the TopicSeg model to ensure that the Bayesian segmentation methodology can achieve reasonable results for segmenting learner essays by topic. The results on the TopicSeg-

Model	Dataset	P_k	WD
TopicSeg	TopicSeg-Tokens	0.213	0.215
TopicSeg	TopicSeg-PTB	0.451	0.497
L1Seg	L1Seg-PTB unigrams	0.473	0.497
L1Seg	L1Seg-PTB bigrams	0.497	0.498
L1Seg	L1Seg-PTB trigrams	0.495	0.490
L1Seg-Compact	L1Seg-PTB bigrams	0.487	0.506
L1Seg-Compact	L1Seg-PTB trigrams	0.392	0.397
L1Seg-Compact	L1Seg-CLAWS2 bigrams	0.389	0.401
L1Seg-Compact	L1Seg-CLAWS2 trigrams	0.373	0.375
L1Seg-AsymP	L1Seg-CLAWS2 trigrams	0.316	0.318
L1Seg-Compact	L1Seg-CLAWS2 trigrams (held-out set)	0.354	0.356
L1Seg-AsymP	L1Seg-CLAWS2 trigrams (held-out set)	0.266	0.271

Table 7.1: Results for all of our text segmentation experiments.

Tokens dataset show that content words are very effective at segmenting the writings by topic, achieving P_k values in the range 0.19–0.21.[20] These values are similar to those reported for segmenting *Wall Street Journal* text (Beeferman et al., 1999). On the other hand, using the PTB POS tag version of the data in the TopicSeg-PTB dataset results in very poor segmentation results, with P_k values around 0.45. This demonstrates that, as expected, POS tags do not provide enough information for topic segmentation; it is not possible to construct even an approximation to lexical chains using them.

7.5.2 L1-based Segmentation

Having verified that the Bayesian segmentation approach is effective for topic segmentation on this data, we now turn to the style-based L1Seg model for segmenting by the native language.

From the results in Table 7.1 we see very poor performance with a P_k value of 0.473 for segmenting the texts in L1Seg-PTB using the unigrams as is. This was a somewhat unexpected result given than we know POS unigram distributions are able to capture syntactic differences between L1-groups — as demonstrated in §3.1.3.5 and §5.5.1 — albeit with limited accuracy.

These results, although not encouraging, are not entirely surprising given that POS tag unigrams are a weak feature relative to other syntactic information such as POS bigrams and trigrams. Accordingly, we considered that the issue could be related to the bag-of-words approach used by the Bayesian topic segmentation model and that it could be improved by enabling the model to take bigrams and trigrams into account, as we described in §7.3.1.3.

[20] These values are obtained by calculating the average segmentation performance across all documents in the dataset, a standard practice in the field.

We repeated the above experiment with both the bigram and trigram encoded versions of L1Seg-PTB, but neither approach resulted in any improvement in our results. In sum, we note that none of the L1Seg variants tested here worked.

7.5.3 Incorporating Discriminative Features

The segmentation results are shown in Table 7.1. Filtering the bigrams results in some minor improvements over the best results from the L1Seg model. However, there are substantial improvements when using the filtered POS trigrams, with a P_k value of 0.392. We did not test unigrams as they were the weakest NLI feature of the three.

This improvement is, we believe, because the Bayesian modelling of lexical cohesion over the input tokens requires that each segment concentrates its probability mass on a compact subset of words. In the context of the n-gram tokenization method tested in the previous section, the L1Seg model with n-grams would most likely exacerbate the issue by prodigiously increasing the number of tokens in the language model. Essentially we are faced with a *feast or famine* scenario: the unigrams do not capture enough information to distinguish non-lexical shifts and the n-grams provide too many features.

We also see that using the CLAWS2 tagset outperforms the PTB tagset. The results achieved for bigrams are much higher, while the trigram results are also better, with $P_k = 0.37$. NLI experiments using different POS tagsets have established that more fine-grained tagsets (*i.e.* those with more tag categories) provide greater classification accuracy when used as n-gram features for classification.[21] These results comport with these previous findings.

In sum, the results from this experiment demonstrate the importance of inducing a compact distribution, which we did here by reducing the vocabulary size by stripping non-informative features.

7.5.4 Applying Two Asymmetric Priors

Our final model, L1Seg-AsymP, assesses whether setting different priors for each L1 can improve performance. We first investigate whether there is any choice of priors that can perform better on our dataset and then verify this on a held-out test set.

The effects of prior strength in Bayesian estimation have been examined, with weak (< 1) and strong (> 1) priors shown to achieve different results. In the context of hierarchical Dirichlet processes, Wang and Blei (2009) note that the value of the prior controls both the smoothing and sparsity of the Dirichlet, where weak priors encourage more sparse and compact models that place their mass only only a few terms . This weak prior is also said to "(encode) increased confidence in the estimate from the observed counts [and] (as) the parameter approaches zero, the expectation of each per-term probability becomes closer to its empirical estimate". An example of how such strong and weak asymmetric priors have been successfully used for sentiment analysis was earlier described in §7.4.4.

We begin with a coarse grid search where the prior search space was defined to be in the interval $[0.1, 3.0]$, partitioned into 30 evenly spaced values that include both weak and strong priors. Values for θ_a and θ_b were selected from these 30 values, resulting in 900 possible prior combinations. These combinations also include cases where θ_a and θ_b both take the same value, making them equivalent to the previous experiment.

[21] See §3.1.3.5 and §5.8 for more details.

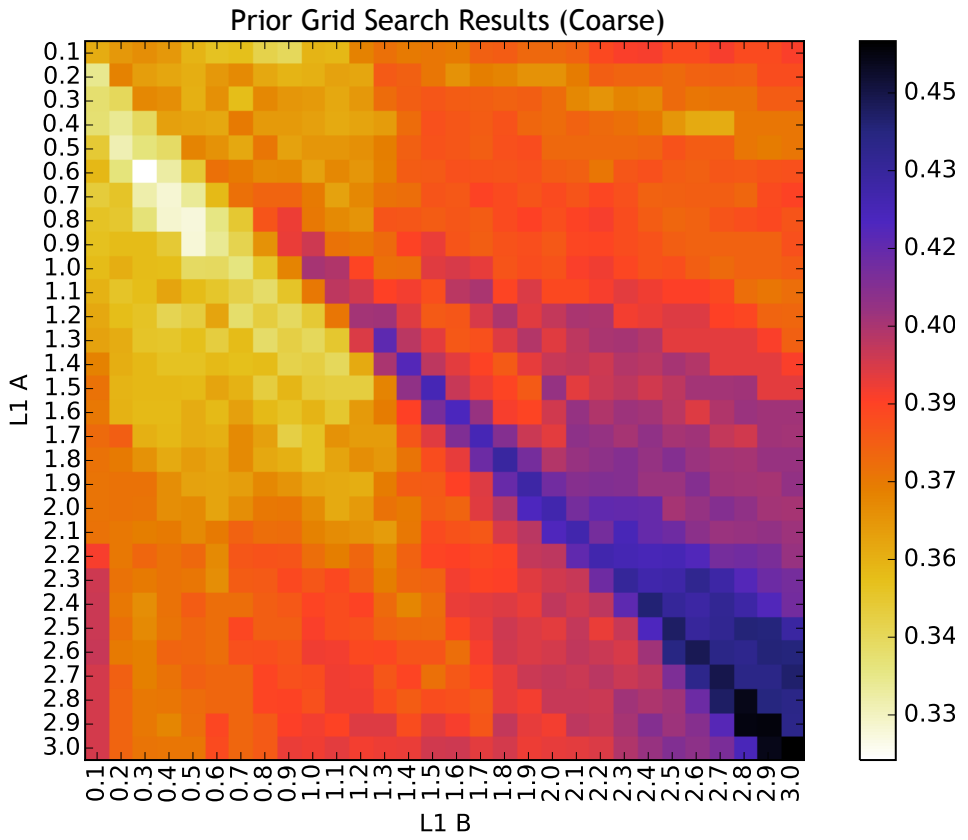

Figure 7.4: Results for running a coarse grid search for an asymmetric prior pair.

The grid search results are visualized in Figure 7.4. We observe that the prior pair of $(0.6, 0.3)$ achieves a P_k value of 0.321, a substantial improvement over the previous best result of 0.37. This also highlights the importance of using weak priors for both values and is consistent with the emphasis on compactness since weak priors result in more compact models.

Next we conduct a more fine-grained grid search, focusing on the range that provided the best results in the coarse search. This time the prior search space was defined to be in the interval $[0.3, 0.9]$, partitioned into 60 evenly spaced values, resulting in 3,600 possible prior combinations.

The results are illustrated in Figure 7.5, showing that the prior pair of $(0.64, 0.32)$ provides a slight improvement of the P_k value to 0.316.

These results thus far have demonstrated that setting an asymmetric prior can be useful for this task. However, we also need to evaluate the prior values on previously unseen data. As the data used in our experiments so far comes from the TOEFL11 test and development sets, we can use the held out TOEFL11-TEST set data for this purpose.

To do this we generate an equivalent of the L1SEG-CLAWS2 dataset, but only using data from the held-out TOEFL11-TEST set, resulting in a total of 5 documents with 5 segments each. Running the L1SEG-COMPACT model on this dataset, we obtain a P_k value of 0.354. This value is very similar to what we obtained on the L1SEG-CLAWS2 data and can be considered a baseline for evaluating our asymmetric prior. Applying the best asymmetric prior from the grid search this improves to 0.266, showing that the grid search was not just a result of overfitting on the training set.

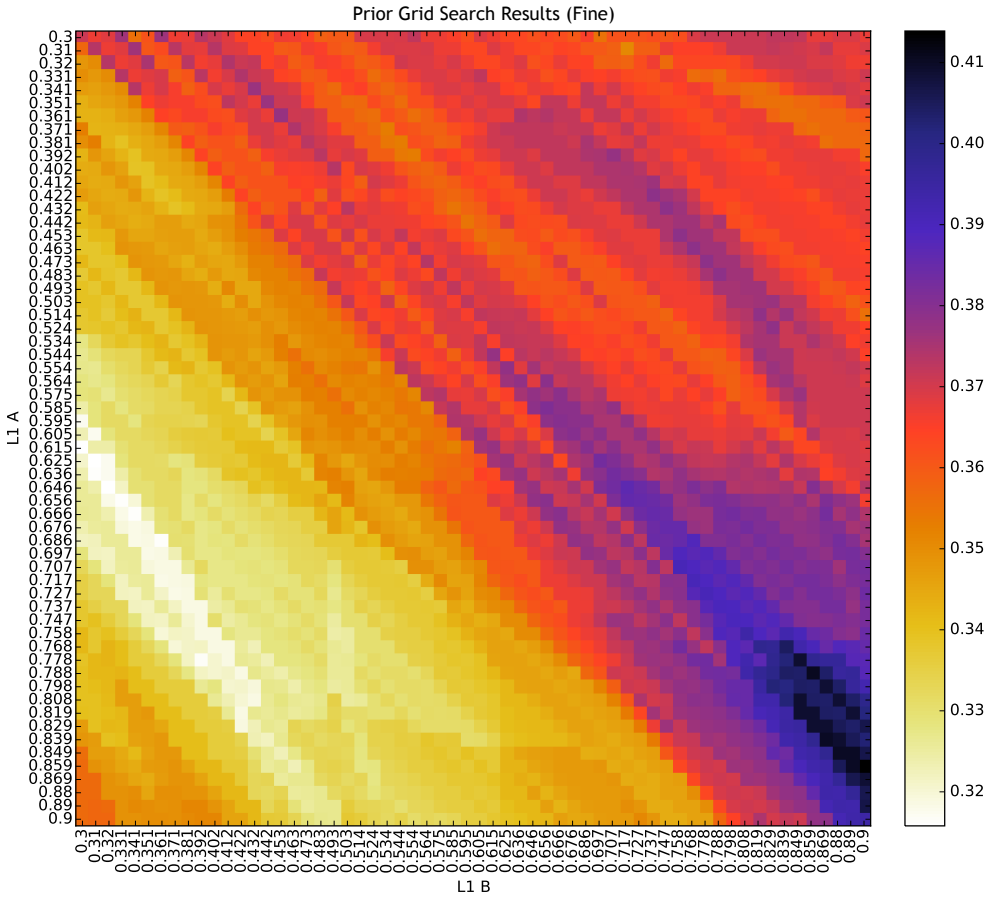

Figure 7.5: Results for running a fine-grained grid search for an asymmetric prior pair.

7.6 Discussion

In this chapter we considered the task of identifying L1 influence in texts that contain such effects from more than one language. We framed this as a variant of text segmentation where the goal is to subdivide a text into regions of different L1 influence. We explored a range of Bayesian models for this, based on adapting an unsupervised topic segmentation approach to use POS n-gram features that capture some short-range syntactic aspects of the text.

We demonstrated that a generative model for stylistic segmentation is possible, and further that segmentation results improve substantially by compacting the n-gram distributions, achieved by incorporating knowledge about discriminative features extracted from NLI models. Our best results come from a model that uses alternating asymmetric priors for each L1, with the priors selected using a grid search and then evaluated on a held-out test set.

This method could be adapted for use in some of the applications we discussed in §7.2, such as literary analysis and plagiarism detection. For intrinsic plagiarism, for example, this could be identified through the detection of stylistic inconsistencies or changes within the text. The detection of such inconsistencies could also be applied to other tasks, such as detecting vandalism on Wikipedia.

Evaluation is another important aspect of this task that requires further consideration. Although we use the same segmentation metrics for topic segmentation and our results here were quite similar to those for topic segmentation, this task is a fundamentally different, and potentially more difficult. The early topic segmentation work described in §7.2.1.1 was initially evaluated against human judgements. We believe that such an approach should also be applied here. While human evaluators have been

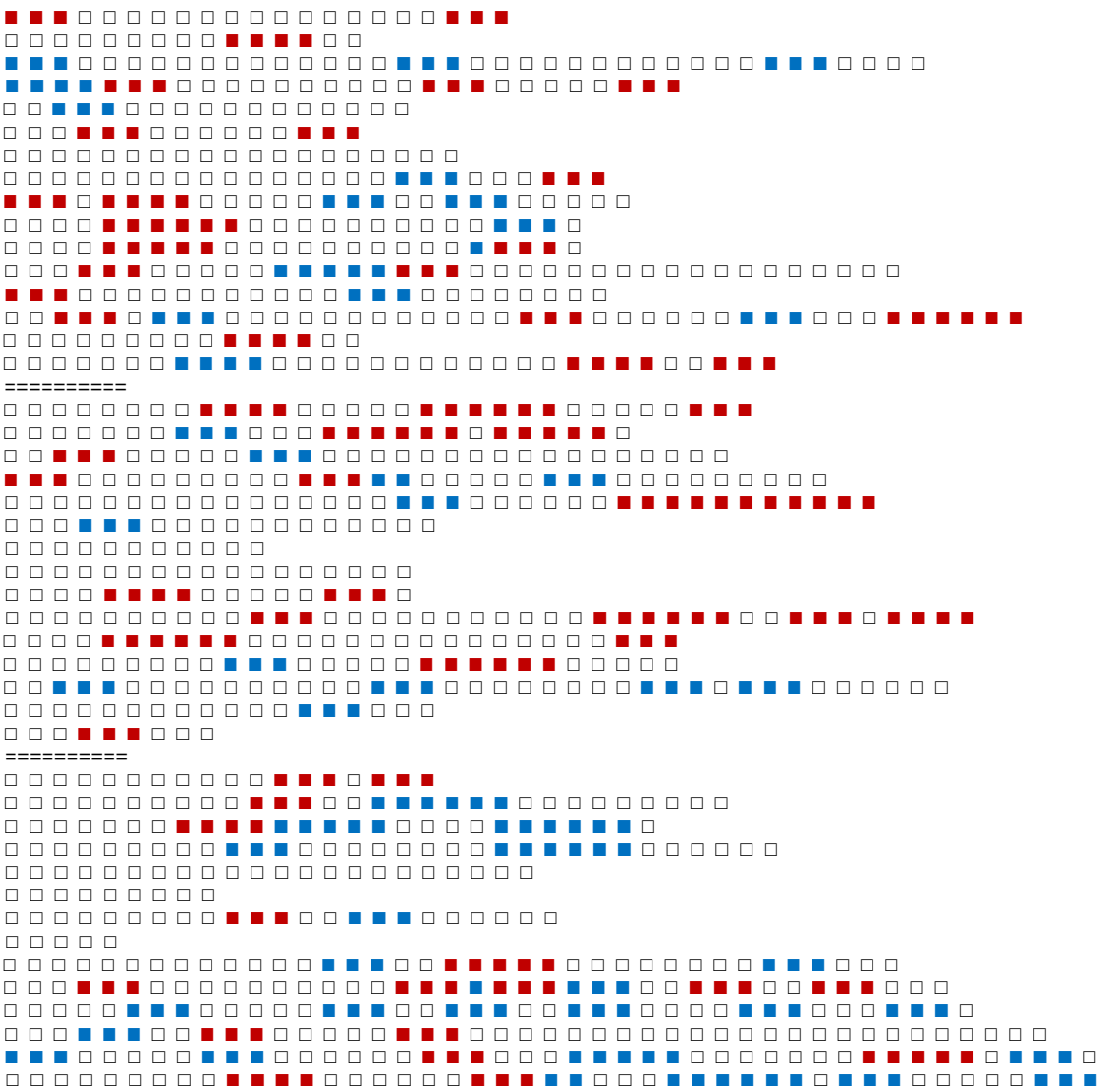

Figure 7.6: A visualization of a document containing three text segments from two alternating L1 backgrounds.

shown to be efficient at detecting topic segments, it is not clear if this would be the case for stylistically-distinct L1-based segments. In §4.1 we showed that human raters could not outperform our NLI system, but it remains to be seen if this pattern is also reflected for the current task. This is left for future work. Perhaps it is a task only specialized literary scholars would be able to attempt.

An examination of one of our documents highlights the difficulty of the task. Figure 7.6 shows a visualization of a document containing three text segments from two alternating L1 backgrounds. Looking at the features of the sentences prior to and following the boundaries, we do not observe any sharp changes in the distribution of the blue and red tokens. This lack of a sharp boundary attests the inherent difficulty of this task. In the absence of the provided boundary positions, it would be difficult to specify their positions with great confidence.

Future work could also consider different types of data. One idea is to investigate using manual annotations of Joseph Conrad's work for literary analysis similar to the scenarios described in §7.1. Yet another idea is to look at plagiarism data, particularly those mixing native and non-native writings. It would be interesting to compare these results against the current segmentation experiment since our

previous work (§5.7) shows that discriminative models can discern non-native writing with extremely high accuracy.

Our previous NLI work has shown that other syntactic features, including CFG production rules and dependency parses, can be useful for capturing L1-based syntactic differences. The incorporation of these features for this segmentation task could also be a potentially fruitful line of inquiry for future work. We have taken a fairly straightforward approach which modifies the generative story. A more sophisticated approach would be to incorporate features into the unsupervised model. One such example is the work of Berg-Kirkpatrick et al. (2010) where they demonstrate that "each component multinomial of a generative model can be turned into a miniature logistic regression model" with the use of a modified EM algorithm. Their results showed that the feature-enhanced models which incorporate linguistically-motivated features achieve substantial improvements for tasks such as part-of-speech induction, grammar induction, word alignment, and word segmentation.

7.7 Chapter Summary

In this chapter we:

- defined the novel task of native language-based text segmentation
- explored a range of Bayesian models for addressing this task
- adapted an unsupervised Bayesian approach originally developed for topic segmentation to use syntactic POS features
- incorporated knowledge about discriminative features extracted using NLI to compact the POS n-gram distribution
- showed that segmentation results improve substantially by compacting the distribution of POS n-grams in this manner
- demonstrated that performance can be further enhanced through extending the segmentation model to incorporate asymmetric priors

Chapter 8

Conclusion

When the research for this thesis commenced in 2012, fewer than 20 papers examining NLI had been published. This was followed by an efflorescence of work in the area and growing interest from researchers. This thesis makes contributions in three broad areas: (1) exploring the task in new ways; (2) investigating how NLI can inform SLA, and (3) introducing the novel task of L1-based text segmentation. In this chapter we briefly recap our major findings and highlight some avenues for future work.

8.1 Summary of Contributions

We began by presenting the development of our NLI system in Chapter 3. We built this system by exploring various classifier ensemble methods, showing that an ensemble of heterogeneous base learners can achieve good results. Analysis of our results showed that larger, more fine-grained POS tagsets are more useful for NLI and that proficiency-segregated models could be useful for improving accuracy. This analysis was expanded with our proposal of a methodology for measuring feature diversity. We highlighted several interesting trends and correlations between lexical and syntactic features, including that syntactic features have lower levels of inter-correlation (and thus higher diversity) compared to their lexical counterparts. In this chapter we also proposed a new feature for NLI: function word n-grams, showing it to be informative for the task. We also demonstrated that word skip-grams can serve as a useful alternative to grammatical dependency features.

In Chapter 4 we considered aspects of NLI that are relevant to practical applications. The first part examined potential reference values for evaluating NLI results and looked at the use of an oracle classifier to estimate a system's performance upper-bound, something which had not been done to date. We also proposed an Accuracy@N measure for evaluating how close a system is to reaching the oracle performance by checking if the true label is within the top N ranked output labels. We created a dataset composed of all the submissions to the 2013 NLI Shared Task, a resource which we made publicly available, and applied our oracle measures to this data. Results here show that for many misclassified texts the correct class label receives a significant portion of the ensemble votes and when not the winning label, it is often the runner-up. This knowledge can also be used to increase NLI accuracy by aiming to develop more sophisticated classifiers that can take into account the top N labels in their decision making, similar to discriminative re-ranking methods applied in statistical parsing. Another question that has been raised concerns how humans would perform at NLI. In the

second part of this chapter we presented the first such study with a group of 10 human experts. While some participants achieved modest results on our simplified setup with 5 L1s, they did not outperform our NLI system, and this performance gap is likely to widen on the standard NLI setup. The difficulty of finding qualified human experts was another issue highlighted here.

The second part of Chapter 4 we considered how well NLI systems might perform on other data, including those from other domains and genres, by conducting large-scale cross-corpus evaluation. Results showed that NLI features may not generalize well across genres and that certain lexical features may exert undue influence on results and leave systems vulnerable to manipulation.

In Chapter 5 we presented the first large-scale study of NLI applied to texts written in languages other than English, using data from six new languages. A cornerstone of this chapter was the identification of suitable data sources which could be used here and for future work. With this goal in mind, we identified six typologically very different sources of non-English L2 data and conducted six experiments using a set of commonly used features. The results of our extensive experiments suggest that NLI accuracy is similar across the set of L2s and a wide range of L1s. We also show that other result patterns, *e.g.* oracle performance and feature diversity, also hold across languages. Another outcome of this work was the release of the Jinan Chinese Learner Corpus (JCLC), containing approximately 6 million Chinese characters written by students from over 50 different L1 backgrounds.

In this chapter we also provided a multilingual assessment of how the degree of syntactic data encoded in part-of-speech tags affects their efficiency as classification features. It was found that while using larger tagsets can lead to improved NLI accuracy, most differences between L1 groups lie in the ordering of the most basic word categories.

In Chapter 6 we shifted our focus towards practical uses of our models in SLA research. Our first exploration here focused on language transfer, the characteristic second language usage patterns caused by native language interference, which is investigated by SLA researchers seeking to find overused and underused features. A methodology for deriving ranked lists of such discriminative features was developed and presented. To illustrate its potential to inform SLA research, we examined some of the most discriminative features and found that they correspond to phenomena and language transfer hypotheses discussed in the SLA literature. This was also the first such experiment to extend NLI to a broad linguistic interpretation of the data and address the automatic extraction of underused features on a per L1 basis.

The second half of Chapter 6 extended this to another research goal in SLA, which is to formulate and test hypotheses about errors and the environments in which they are made, a process which often involves substantial effort and large amounts of data; computational techniques promise help here. To this end we defined a new task for finding contexts for errors that vary with the native language of the speaker that are potentially useful for SLA research. We propose four models for approaching this task, and find that one based only on error-feature co-occurrence and another based on determining maximum weight cliques in a feature association graph discover strongly distinguishing contexts, with an apparent trade-off between false positives and very specific contexts.

The final experiments of this thesis – described in Chapter 7 – explore the application of non-lexical language transfer features to the novel task of native language-based text segmentation. The aim of this task is to subdivide a text into regions that exhibit differing L1 influence. Such methods could be applied for intrinsic plagiarism detection or even literary analysis. We adapted an unsupervised Bayesian approach originally developed for topic segmentation to use syntactic POS features.

Segmentation results were substantially improved by compacting the distribution of POS n-grams, achieved by selective inclusion of the most discriminative features obtained using the methodology we presented in Chapter 6. We also demonstrated that performance can be further enhanced through extending the segmentation model to incorporate asymmetric priors and empirical selection of these priors using a grid search.

Resource Contributions Research from this thesis has also resulted in the development of two publicly available resource contributions:

- **The NLI2013 Submissions dataset** — containing the 144 entries submitted to 2013 NLI shared task, as described in §4.1.4. This is useful for both NLI and ensemble research.
- **The Jinan Chinese Learner Corpus** — a dataset of approximately 6 million Chinese characters written by students from over 50 different L1 backgrounds, as described in §5.2.4.
- **BayesSeg with Asymmetric Priors** — we are working on releasing the code for our extension to BayesSeg for supporting asymmetric priors, as described in §7.4.4.

8.2 Future Work

Given the multi-faceted nature of the work presented here, there are many directions for future work. Many of these potential avenues were detailed within each chapter and experiment presented here; in most cases these ideas were local and mostly related to the methodological aspects of each experiment. We highlight a select few recommendations for future expansions of the experiments here.

The classification systems presented in Chapter 3 could benefit from the application of more sophisticated ensemble methods, *e.g.* ensemble pruning (Zhang et al., 2006) or stacked generalization (Ting and Witten, 1999), to improve NLI accuracy. The commission of a larger study of human NLI performance like the one we presented in Chapter 4 can also be a valuable contribution that could lead to further insights, although this would likely be a costly and time-consuming endeavour. Wider adoption of oracles in future NLI experiments can help better contextualize the results and assist with their interpretation. Our multilingual NLI experiments from Chapter 5 could also be expanded to additional languages, although this is limited by the current paucity of suitable data. Following on from our experiments on discriminating the writings of native speakers from non-natives, one direction for expansion of the NLI methodology is the inclusion of a Native class within experiments. In the contexts of work with the TOEFL11 data, for example, this would involve the inclusion of an English L1 class.

We also pointed to the failure to distinguish between the L2 and any other acquired languages as a more general criticism of the NLI literature to date. The current body of NLI literature fails to distinguish whether the learner language is in fact the writer's second language, or whether it is possibly a third language (L3). None of the corpora used in this thesis contain this metadata. It has been noted in the SLA literature that when acquiring an L3, there may be instances of both L1- and L2-based transfer effects on L3 production (Ringbom, 2001). Studies of such second language transfer effects during L3 acquisition have been a recent focus in cross-linguistic influence research (Murphy, 2005).

Although one very promising application of NLI is in the forensic domain, not much effort has been expended in specifically assessing the methods for such tasks. Such a shift would require data that goes beyond learner essays and instead more closely resembles real word scenarios, *e.g.* anonymous letters, threats and radical communications. It would also be interesting to see to what degree these methods are susceptible to wilful deception on the part of the authors.

In addition to the specific propositions listed so far, there are at least two distinct branches that can be followed for long term research planning. The first of these concerns how we can utilize the information that our models have learnt about different L1 groups. In particular, this relates to how this information can be applied towards SLA research and language teaching. Making the connection between these areas will require that the information in our models be presented in a meaningful way that is conducive to the task at hand. For example, from a pedagogical viewpoint, knowledge derived here could assist with language assessment and curriculum design. This research also holds out the promise of identifying previously hidden issues in non-native writing, and allow language transfer hypotheses to be formulated and reliably tested on a tremendously larger scale than previously possible.

A second branch stems from the type of work we presented in Chapter 7 and is concerned with how the computational models generated for NLI can be exploited for other NLP tasks. This is particularly true for applications where knowledge about the L1 is helpful. The related tasks of grammatical error detection and correction are salient examples of tasks which are often applied to non-native writing where L1-specific knowledge could be applied to build more accurate models that take into account native language-based idiosyncrasies of the input. Automated essay grading is another related task knowledge from our models could be applied.

Methods for the detection of non-native writing could also be applied to tasks such as machine translation in order to select the most appropriate methods and models based on the input. Similarly, machine translation of minor languages still relies on significant post-editing work and in cases of human translations that are not fluent, models similar to our work can help identify errors and ungrammatical fragments.

Automated Speech Recognition (ASR) is another NLP task where non-native language production poses serious issues. While there has been work in creating models that can adapt to non-native pronunciation (Bouselmi et al., 2006), most of this work in non-native speech decoding has focused on acoustic models. One approach here has been the addition of potential L1-induced mispronunciations that non-natives may produce (Goronzy et al., 2004; Livescu and Glass, 2000). However, other methods are needed to compensate for deviations in syntax (Van Doremalen et al., 2010). This also includes ASR-based Computer-Assisted Language Learning (CALL) systems that need to provide intelligent feedback on aspects of a speaker's performance such as pronunciation and syntax.

In this context, the L1-specific linguistic features captured through NLI could be integrated within such ASR systems in order to detect deviations from native speech patterns. As most ASR systems rely on *maximum a-posteriori* approximation, accounting for syntactic deviations in word order and sentence structure could better enable the decoder to map the acoustic observations to the actual, albeit erroneous, sequence of words produced by the speaker. Going further, such systems could then use this information to perform error correction as well.

Bibliography

Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group Web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005. 7, 19, 110

May Ali Abdul-Ameer and Khalid Shakir Hussein. Plagiarism and patchwriting detection in efl students' graduation research writing. *Research on Humanities and Social Sciences*, 5(8):128–136, 2015. 168

Amjad Abu-Jbara, Rahul Jha, Eric Morley, and Dragomir Radev. Experimental Results on the Native Language Identification Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–88, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1710`. 42

Ghazi Abuhakema, Reem Faraj, Anna Feldman, and Eileen Fitzpatrick. Annotating an Arabic Learner Corpus for Error. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2004)*, 2008. 104

Charles S. Ahn. Automatically Detecting Authors' Native Language. Master's thesis, Naval Postgraduate School, Monterey, CA, 2011. 37

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9(11):1–12, 2008. 92

Abdullah Alfaifi and Eric Atwell. Arabic Learner Corpus v1: A New Resource for Arabic Language Research. In *Proceedings of the Second Workshop on Arabic Corpus Linguistics*, 2013. 110

Abdullah Alfaifi, Eric Atwell, and I Hedaya. Arabic learner corpus (ALC) v2: a new written and spoken corpus of Arabic learners. In *Proceedings of the Learner Corpus Studies in Asia and the World (LCSAW)*, Kobe, Japan, 2014. URL `http://www.arabiclearnercorpus.com/`. 110

Josh Ard and Taco Homburg. Verification of language transfer. In Susan M Gass and Larry Selinker, editors, *Language transfer in language learning*, pages 157–176. Newbury House MA, 1983. 52

Shlomo Argamon, Moshe Koppel, James Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2006. 20

Terry Kit-Fong Au. Chinese and English counterfactuals: the Sapir-Whorf hypothesis revisited. *Cognition*, 15(1):155–187, 1983. 143

Roger Bakeman and Vicenç Quera. *Sequential analysis and observational methods for the behavioral sciences*. Cambridge University Press, 2011. 74

Peter Ball. Stereotypes of anglo-saxon and non-anglo-saxon accents: Some exploratory australian studies with the matched guise technique. *Language sciences*, 5(2):163–183, 1983. 5

Calvin Bedient and Thomas Stearns Eliot. *He Do the Police in Different Voices: The Waste Land and its protagonist*. University of Chicago Press, 1986. 167

Doug Beeferman, Adam Berger, and John Lafferty. Text Segmentation Using Exponential Models. In *Second Conference on Empirical Methods in Natural Language Processing*, pages 35–46, 1997. URL `http://www.aclweb.org/anthology/W97-0304`. 163

Doug Beeferman, Adam Berger, and John Lafferty. Statistical Models for Text Segmentation. *Machine Learning*, 34(1-3):177–210, February 1999. ISSN 0885-6125. doi: 10.1023/A:1007506220214. URL `http://dx.doi.org/10.1023/A:1007506220214`. 163, 172, 178

Kenneth R Beesley. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, volume 47, page 54, 1988. 16

Nicholas J. Belkin and W. Bruce Croft. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Communications of the ACM*, 35(12):29–38, December 1992. URL `http://doi.acm.org/10.1145/138859.138861`. 17

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590. Association for Computational Linguistics, 2010. 183

Shane Bergsma, Matt Post, and David Yarowsky. Stylometric Analysis of Scientific Articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, Montréal, Canada, June 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N12-1033`. 20

Derek Bickerton. *Roots of language*. Karoma, 1981. 135

Laada Bilaniuk. Gender, language attitudes, and language status in Ukraine. *Language in Society*, 32(01):47–78, 2003. 5

Susan Blackwell. *History of Forensic Linguistics*. Blackwell Publishing Ltd, 2013. ISBN 9781405198431. doi: 10.1002/9781405198431.wbeal0508. URL `http://dx.doi.org/10.1002/9781405198431.wbeal0508`. 6, 7

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. TOEFL11: A Corpus of Non-Native English. Technical report, Educational Testing Service, 2013. 25, 26, 28, 104

David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, April 2012. ISSN 0001-0782. 173

Alfred H Bloom. *The linguistic shaping of thought: A study in the impact of language on thinking in China and the West*. Psychology Press, 2014. 143

Victoria Bobicev. Native Language Identification with PPM. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 180–187, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1724`. 42

I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo. The Maximum Clique Problem. In D.-Z. Du and P. M. Pardalos, editors, *Handbook of Combinatorial Optimization (supp. Vol. A)*, pages 1–74. Kluwer Academic, Dordrecht, Netherlands, 1999. 152

Ghazi Bouselmi, Dominique Fohr, Irina Illina, and Jean-Paul Haton. Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints. In *The Ninth International Conference on Spoken Language Processing-ICSLP 2006*, 2006. 188

Michael Branch. Finnish. In Bernard Comrie, editor, *The world's major languages*, pages 497–518. Routledge, 2009. 111

Leo Breiman. Bagging Predictors. *Machine Learning*, pages 123–140, 1996. 57, 65

Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language*, pages 112–116. Association for Computational Linguistics, 1992. 32

Ted Briscoe, John Carroll, and Rebecca Watson. The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 77–80, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. 150

Ted Briscoe, Ben Medlock, and Øistein Andersen. Automated Assessment of ESOL Free Text Examinations. Technical Report TR-790, University of Cambridge, Computer Laboratory, 2010. 147

Julian Brooke and Graeme Hirst. Native language detection with 'cheap' learner corpora. In *Conference of Learner Corpus Research (LCR2011)*, Louvain-la-Neuve, Belgium, 2011. Presses universitaires de Louvain. 23, 37, 92, 93, 109, 118

Julian Brooke and Graeme Hirst. Measuring interlanguage: Native language identification with L1-influence metrics. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 779–784, Istanbul, Turkey, May 2012a. 37

Julian Brooke and Graeme Hirst. Robust, Lexicalized Native Language Identification. In *Proceedings of COLING 2012*, pages 391–408, Mumbai, India, December 2012b. The COLING 2012 Organizing Committee. URL `http://www.aclweb.org/anthology/C12-1025`. 23, 43, 56, 60, 72, 115, 116, 130

Julian Brooke and Graeme Hirst. Using Other Learner Corpora in the 2013 NLI Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 188–196, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1725`. 43, 47, 70

Julian Brooke, Adam Hammond, and Graeme Hirst. Unsupervised Stylistic Segmentation of Poetry with Change Curves and Extrinsic Features. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 26–35, Montréal, Canada, June 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W12-2504`. 167, 169, 174

Julian Brooke, Graeme Hirst, and Adam Hammond. Clustering voices in the waste land. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 41–46, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1406`. 167

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479, 1992. 46

Christopher JC Burges. A tutorial on Support Vector Machines for Pattern Recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998. 140

Serhiy Bykh and Detmar Meurers. Native Language Identification using Recurring n-grams – Investigating Abstraction and Domain Dependence. In *Proceedings of COLING 2012*, pages 425–440, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL `http://www.aclweb.org/anthology/C12-1027`. 38, 56

Serhiy Bykh and Detmar Meurers. Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1962–1973, August 2014. 34, 48, 77, 92

Serhiy Bykh, Sowmya Vajjala, Julia Krivanek, and Detmar Meurers. Combining Shallow and Linguistically Motivated Features in Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 197–206, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1726`. 41

Aoife Cahill. Parsing Learner Text: to Shoehorn or not to Shoehorn. In *The 9th Linguistic Annotation Workshop*, page 144, 2015. 72

Bert Cappelle and Rudy Loock. Is there interference of usage constraints?: A frequency study of existential there is and its French equivalent il ya in translated vs. non-translated texts. *Target*, 25 (2):252–275, 2013. 144

Xavier Carreras, Isaac Chao, Lluis Padró, and Muntsa Padró. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 2004. 114

William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US, 1994. 16, 29

Jasone Cenoz and Ulrike Jessner. *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*, volume 31. Multilingual Matters, 2001. 131, 203

Yin-Wen Chang and Chih-Jen Lin. Feature ranking using linear SVM. *Causation and Prediction Challenge Challenges in Machine Learning, Volume 2*, page 47, 2008. 139

Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180. Association for Computational Linguistics, 2005. 36, 85

Carole E. Chaski. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1):1–13, 2005. 19

Harr Chen, SRK Branavan, Regina Barzilay, David R Karger, et al. Content modeling using latent permutations. *Journal of Artificial Intelligence Research*, 36(1):129–163, 2009. 166

Jianguo Chen, Chuang Wang, and Jinfa Cai. *Teaching and learning Chinese: Issues and perspectives*. IAP, 2010. 102

Meilin Chen. Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics*, 18(3), 2013. 138

Kenneth Church. A Pendulum Swung Too Far. *Linguistic Issues in Language Technology*, 6(5), 2011. 10

Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. Linguistic Profiling based on General–purpose Features and Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 207–215, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1727`. 41

A. P. Coleman. Polonisms in the English of Conrad's Chance. *Modern Language Notes*, 46(7):pp. 463–468, 1931. ISSN 01496611. URL `http://www.jstor.org/stable/2913489`. 160

Bernard Comrie. *Language Universals and Linguistic Typology*. University of Chicago Press, Chicago, IL, US, 2nd edition, 1989. 110

John Xiros Cooper. *TS Eliot and the politics of voice: The argument of The Waste Land*. Number 79. UMI Research Press, 1987. 167

Stephen P. Corder. The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 5(4):161–170, 1967. 49

Elisa Corino. VALICO: An Online Corpus of Learning Varieties of the Italian Language. In *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics*, pages 117–133, 2008. 105

Malcolm Coulthard and Alison Johnson. *An introduction to Forensic Linguistics: Language in evidence*. Routledge, 2007. 6

Vidas Daudaravicius. VTEX System Description for the NLI 2013 Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 89–95, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1711`. 42

Gessica De Angelis. Multilingualism and non-native lexical transfer: An identification problem. *International Journal of Multilingualism*, 2(1):1–25, 2005. 131

Annette De Groot and Rik Poot. Word translation at three levels of proficiency in a second language: The ubiquitous involvement of conceptual memory. *Language learning*, 47(2):215–264, 1997. 52

Marie-Catherine de Marneffe, Bill Maccartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 449–454, Genoa, Italy, 2006. 61

Gerard de Melo. Etymological Wordnet: Tracing The History of Words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA), 2014. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/1083_Paper.pdf`. 79

Jean-Marc Dewaele. Lexical inventions: French interlanguage as L2 versus L3. *Applied Linguistics*, 19(4):471–490, 1998. 131

Mona Diab. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*, 2009. 110

Ana Díaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1-2):139–154, 2010. 72

Markus Dickinson and W. Detmar Meurers. Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 107–114, Budapest, Hungary, 2003. 39, 56

Markus Dickinson and W. Detmar Meurers. Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 322–329, 2005. 39

María Belén Diéz-Bedmar and Szilvia Papp. The use of the English article system by Chinese and Spanish learners. *Language and Computers*, 66(1):147–176, 2008. 2, 134, 146, 150, 153

Lan Du, Wray L Buntine, and Mark Johnson. Topic Segmentation with a Structured Topic Model. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2013)*, pages 190–200, 2013. 166

Jacob Eisenstein. Hierarchical text segmentation from multi-scale lexical cohesion. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 353–361, Boulder, CO, 2009. URL `www.aclweb.org/anthology/N09-1040`. 165, 175

Jacob Eisenstein and Regina Barzilay. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D08-1035`. 159, 162, 164, 165, 166, 170, 172, 173, 175, 176

Miriam Eisenstein. Native reactions to non-native speech: A review of empirical research. *Studies in Second Language Acquisition*, 5(2):160–176, 1983. 5

Rod Ellis. *The study of second language acquisition*. Oxford University Press, 1994. 49

Rod Ellis. *The Study of Second Language Acquisition*. Oxford University Press, Oxford, UK, 2nd edition, 2008. 2, 51, 136

Gerard Escudero, Lluís Màrquez, and German Rigau. An empirical study of the domain dependence of supervised word sense disambiguation systems. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 172–180. Association for Computational Linguistics, 2000. 16, 92

Dominique Estival, Tanja Gaustad, Son-Bao Pham, Will Radford, and Ben Hutchinson. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 263–272, Melbourne, Australia, September 2007. 20, 35

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008. 60

George Forman. A pitfall and solution in multi-class feature selection for text classification. In *Proceedings of the twenty-first international conference on Machine learning*, page 38. ACM, 2004. 139

Yoav Freund and Robert E Schapire. Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 96, pages 148–156, 1996. 57

Michael Gamon, Jianfeng Gao, Chris Brockett, Alex Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP'08)*, pages 449–456, Hyderabad, India, 2008. 8

Michael Gamon, Martin Chodorow, Claudia Leacock, and J Tetreault. Using learner corpora for automatic error detection and correction. In N Ballier, A Díaz-Negrillo, and P Thompson, editors, *Automatic treatment and analysis of learner corpus data*, Studies in Corpus Linguistics, pages 127–150. John Benjamins Publishing Company, Amsterdam, The Netherlands, 2013. 22

Lisa Garbern Liu. Reasoning counterfactually in Chinese: Are there any obstacles? *Cognition*, 21 (3):239–270, 1985. 143

Roger Garside. The CLAWS word-tagging system. In *The computational analysis of English: A Corpus Based Approach*. Longman, London, 1987. 125

Susan M. Gass and Larry Selinker. *Second Language Acquisition: An Introductory Course*. Routledge, New York, 2008. 51, 134

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. Improving Native Language Identification with TF-IDF Weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 216–223, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1728`. 44

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In *Proceedings of the 31st Second Language Research Forum*, 2013. 93

Alexander Genkin, David D Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007. 17

John Gibbons. *Forensic Linguistics: An Introduction To Language In The Justice System*. Wiley-Blackwell, 2003. 6

John Gibbons and Venn Prakasam. *Language in the Law*. Orient Blackswan, 2004. 6

Daniel Gildea. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 167–202, 2001. 92

Howard Giles and Tania Ogay. Communication accommodation theory. *Explaining communication: Contemporary theories and exemplars*, pages 293–310, 2007. 5

Howard Giles, Pamela Wilson, and Anthony Conway. Accent and lexical diversity as determinants of impression formation and perceived employment suitability. *Language Sciences*, 3(1):91–103, 1981. 5

Howard Giles, Justine Coupland, and Nikolas Coupland. *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press, 1991. 5

Silke Goronzy, Stefan Rapp, and Ralf Kompe. Generating non-native pronunciation variants for lexicon adaptation. *Speech Communication*, 42(1):109–123, 2004. 188

Cyril Goutte, Serge Léger, and Marine Carpuat. Feature Space Selection and Combination for Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 96–100, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1712`. 41

Sylviane Granger. The learner corpus: A Revolution in Applied Linguistics. *English Today*, 10(03): 25–33, 1994. 4, 22

Sylviane Granger. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, 37(3):538–546, 2003. 23, 104

Sylviane Granger. How to Use Foreign and Second Language Learner Corpora. In Alison Mackey and Susan M. Gass, editors, *Research Methods in Second Language Acquisition: A Practical Guide*. Wiley-Blackwell, 2011. 2

Sylviane Granger and Stephanie Tyson. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, 15(1):17–27, 1996. 50

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. *International Corpus of Learner English (Version 2)*. Presses Universitaires de Louvain, Louvian-la-Neuve, Belgium, 2009. 35

Tim Grant. Quantifying evidence in forensic authorship analysis. *International Journal of Speech Language and the Law*, 14(1):1–25, 2007. 7

Tim Grant. Text messaging forensics. *The Routledge Handbook of Forensic Linguistics*, page 508, 2010. 7

Tim Grant. TXT 4N6: Method, Consistency, and Distinctiveness in the Analysis of SMS Text Messages. *Journal of Law and Policy*, 21:467, 2012. 7

John N. Green. Spanish. In Bernard Comrie, editor, *The world's major languages*, pages 197–216. Routledge, 2009. 107

Barbara B Greene and Gerald M Rubin. Automated grammatical tagging of English. 1971. 125

Nicholas Groom. Closed-class keywords and corpus-driven discourse analysis. *Keyness in Texts. Amsterdam & Philadelphia: John Benjamins*, pages 59–78, 2010. 54

Yan Guo and Gulbahar H Beckett. The Hegemony of English as a Global Language: Reclaiming Local Knowledge and Culture in China. *Convergence*, 40:117–132, 2007. 102

David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. A Close Look at Skip-gram Modelling. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1222–1225, Genoa, Italy, 2006. 62, 78

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003. 139

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002. 97, 139, 140

Binod Gyawali, Gabriela Ramirez, and Thamar Solorio. Native Language Identification: a Simple n-gram Based Approach. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–231, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1729`. 40, 125

Nizar Y Habash. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187, 2010. 102

M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman Publishing Group, May 1976. 162

Petra Hammer and Madeleine Monod. *English-French Cognate Dictionary*. ERIC, 1976. 53

John A. Hawkins. German. In Bernard Comrie, editor, *The world's major languages*, pages 86–109. Routledge, 2009. 105

Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, USA, June 1994. Association for Computational Linguistics. doi: 10.3115/981732.981734. URL `http://www.aclweb.org/anthology/P94-1002`. 163, 167

Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Lingustics*, 23(1):33–64, 1997. URL `http://www.aclweb.org/anthology/J97-1003`. 163, 167

John Henderson, Guido Zarrella, Craig Pfeifer, and John D. Burger. Discriminating Non-Native English with 350 Words. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 101–110, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1713`. 41, 78

Yves Hervouet. *The French Face of Joseph Conrad*. Cambridge University Press, 1990. 161

Barbora Hladka, Martin Holub, and Vincent Kriz. Feature Engineering in the NLI Shared Task 2013: Charles University Submission Report. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 232–241, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1730`. 41, 78

Tin Kam Ho, Jonathan J. Hull, and Sargur N. Srihari. Decision combination in multiple classifier systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(1):66–75, 1994. 60

David I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998. 19

Marlise Horst, Joanna White, and Philippa Bell. First and second language knowledge in the language classroom. *International Journal of bilingualism*, 2010. 137

EH Hubbard. Errors in court: A forensic application of error analysis. *South African Journal of Linguistics*, 12(sup20):3–16, 1994. 7

Oliver A. Iggesen. *Number of Cases*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `http://wals.info/chapter/49`. 110

Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. Can characters reveal your native language? A language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014. Association for Computational Linguistics. 48, 74

Tania Ionin and Silvina Montrul. The role of L1 transfer in the interpretation of articles with definite plurals. *Language Learning*, 60(4):877–925, 2010. 153

Ilmari Ivaska. The corpus of advanced learner Finnish (LAS2): database and toolkit to study academic learner Finnish. *Apples*, 8, 2014. 111

Sherri L. Jackson. *Statistics: Plain and Simple*. Wadsworth, Cengage Learning, Belmont, CA, US, 2009. 150

Scott Jarvis. The Detection-Based Approach: An Overview. In Scott Jarvis and Scott A. Crosley, editors, *Approaching Language Transfer through Text Classification*, pages 1–33. Multilingual Matters, 2012. 6, 53

Scott Jarvis and Scott Crossley, editors. *Approaching Language Transfer Through Text Classification: Explorations in the Detection-based Approach*. Multilingual Matters, Bristol, UK, 2012. 10, 48, 136, 137

Scott Jarvis, Gabriela Castaneda-Jiménez, and Rasmus Nielsen. Investigating L1 lexical transfer through learners' wordprints. In *Second Language Research Forum (SLRF)*, 2004. 35

Scott Jarvis, Yves Bestgen, and Steve Pepper. Maximizing Classification Accuracy in Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1714`. 42, 74, 84, 115

Minwoo Jeong and Ivan Titov. Unsupervised discourse segmentation of documents with inherently parallel structure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 151–155. Association for Computational Linguistics, 2010. 166

Thorsten Joachims. *Text categorization with Support Vector Machines: Learning with many relevant features*. Springer, 1998. 17

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 99, pages 200–209, 1999. 17

Mark Johnson. PCFGs, Topic Models, Adaptor Grammars and Learning Topical Collocations and the Structure of Proper Names. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1157, Uppsala, Sweden, July 2010. Association for Computational Linguistics. 34

Kees Jong, Elena Marchiori, Michele Sebag, and Aad Van Der Vaart. Feature selection in proteomic pattern data with support vector machines. In *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB'04. Proceedings of the 2004 IEEE Symposium on*, pages 41–48. IEEE, 2004. 140

Alan S. Kaye. Arabic. In Bernard Comrie, editor, *The world's major languages*, pages 560–577. Routledge, 2009. 110

Anna Kazantseva and Stan Szpakowicz. Linear text segmentation using affinity propagation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 284–293. Association for Computational Linguistics, 2011. 166

Casey Keck. The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing*, 15(4):261–278, 2006. 168

Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38. Association for Computational Linguistics, 1997. 16

Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998. 58

Youngjoong Ko and Jungyun Seo. Automatic text categorization by unsupervised learning. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 453–459. Association for Computational Linguistics, 2000. 14

Ekaterina Kochmar. Identification of a writer's native language by error analysis. Master's thesis, Computer Laboratory, St. John's College, University of Cambridge, 2011. 37, 56, 136

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145, 1995. 113

Moshe Koppel and Jonathan Schler. Authorship verification as a one-class classification problem. In *Proceedings of the Twenty-first International Conference on Machine Learning*, Proceedings of the International Conference on Machine Learning (ICML 2004), page 62, New York, NY, USA, 2004. ACM. ISBN 1-58113-838-5. 124

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*, volume 3495 of *Lecture Notes in Computer Science*, pages 209–217. Springer-Verlag, 2005a. 31, 35, 36, 136

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628, Chicago, IL, 2005b. ACM. 35

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26, 2009. 23, 56

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 1356–1364. Association for Computational Linguistics, 2011a. 166

Moshe Koppel, Shlomo Argamon, and Jonathan Schler. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011b. 20

Ludmila I Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on pattern analysis and machine intelligence*, 24(2):281–286, 2002. 82

Ludmila I Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2004. 58

Ludmila I Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, second edition, 2014. 58, 59

Ludmila I Kuncheva and Juan J Rodríguez. A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems*, 38(2):259–275, 2014. 57

Ludmila I Kuncheva, James C Bezdek, and Robert PW Duin. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2):299–314, 2001. 82

Ludmila I Kuncheva, Christopher J Whitaker, Catherine A Shipp, and Robert PW Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications*, 6(1):22–31, 2003. 73, 82

Kristopher Kyle, Scott Crossley, Jianmin Dai, and Danielle McNamara. Native Language Identification: A Key N-gram Category Approach. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 242–250, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1731`. 45, 47, 70

Robert Lado. *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. University of Michigan Press, Ann Arbor, MI, US, 1957. 2, 49, 56, 136

Shibamouli Lahiri and Rada Mihalcea. Using N-gram and Word Network Features for Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 251–259, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1732`. 45, 115

Louisa Lam. Classifier combinations: implementations and theoretical issues. In *Multiple classifier systems*, pages 77–86. Springer, 2000. 73

Wallace E Lambert. A social psychology of bilingualism. *Journal of social issues*, 23(2):91–109, 1967. 5

Leah S Larkey. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 90–95. ACM, 1998. 17

Leah S Larkey and W Bruce Croft. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 289–297. ACM, 1996. 85

Jey Han Lau, Timothy Baldwin, and David Newman. On Collocations and Topic Models. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(3), 2013. 172

Batia Laufer and Nany Girsai. Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, 29(4):694–716, 2008. 137

Thomas Lavergne, Gabriel Illouz, Aurélien Max, and Ryo Nagata. LIMSI's participation to the 2013 shared task on Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 260–265, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1733`. 44

John Sie Yuen Lee. *Automatic correction of grammatical errors in non-native English text*. PhD thesis, Massachusetts Institute of Technology, 2009. 8

Yong-Bae Lee and Sung Hyon Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150. ACM, 2002. 16

Roger Levy and Christopher Manning. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 439–446. Association for Computational Linguistics, 2003. 108

Baoli Li. Recognizing English Learners Native Language from Their Writings. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 119–123, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1715`. 42, 47

Charles N. Li and Sandra A. Thompson. Chinese. In Bernard Comrie, editor, *The world's major languages*, pages 703–723. Routledge, 2009. 108

Patsy M Lightbown. Great expectations: Second-language acquisition research and classroom teaching. *Applied Linguistics*, 6(2):173–189, 1985. 134

Patsy M Lightbown. Anniversary article. Classroom SLA research and second language teaching. *Applied linguistics*, 21(4):431–462, 2000. 138

Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 375–384, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. URL `http://doi.acm.org/10.1145/1645953.1646003`. 175

Chenghua Lin, Yulan He, Richard Everson, and Stefan Rüger. Weakly supervised joint sentiment-topic detection from text. *Knowledge and Data Engineering, IEEE Transactions on*, 24(6):1134–1145, 2012. 175

Haitao Liu. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191, 2008. 77, 78

Huan Liu and Hiroshi Motoda. *Computational methods of feature selection*. CRC Press, 2007. 139

Karen Livescu and James Glass. Lexical modeling of non-native speech for automatic speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, volume 3, pages 1683–1686. IEEE, 2000. 188

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444, 2002. 46

Cristóbal Lozano. CEDEL2: Corpus escrito del Español L2. In Carmen M. et al. Bretones Callejas, editor, *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*, pages 197–212. Universidad de Almería, Almería, 2009. 111

Cristobal Lozanó and Amaya Mendikoetxea. Interface conditions on postverbal subjects: A corpus study of L2 English. *Bilingualism: Language and Cognition*, 13(4):475–497, 2010. 138

Cristóbal Lozano and Amaya Mendikoetxea. Learner corpora and Second Language Acquisition: The design and collection of CEDEL2. *Automatic Treatment and Analysis of Learner Corpus Data. Amsterdam: John Benjamins*, 2013. 104, 111

André Lynum. Native language identification using large scale lexical features. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 266–269, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1734`. 44

Shervin Malmasi and Aoife Cahill. Measuring Feature Diversity in Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 49–55, Denver, Colorado, June 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-0606`.

Shervin Malmasi and Mark Dras. Arabic Native Language Identification. In *Proceedings of the Arabic Natural Language Processing Workshop (EMNLP 2014)*, pages 180–186, Doha, Qatar, October 2014a. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W14-3625`.

Shervin Malmasi and Mark Dras. Chinese Native Language Identification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-14)*, pages 95–99, Gothenburg, Sweden, April 2014b. Association for Computational Linguistics. URL `http://aclweb.org/anthology/E14-4019`. 127

Shervin Malmasi and Mark Dras. Finnish Native Language Identification. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 139–144, Melbourne, Australia, November 2014c. URL `http://www.aclweb.org/anthology/U14-1020`.

Shervin Malmasi and Mark Dras. Language Transfer Hypotheses with Linear SVM Weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1385–1390, Doha, Qatar, October 2014d. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D14-1144`.

Shervin Malmasi and Mark Dras. A Data-driven Approach to Studying Given Names and their Gender and Ethnicity Associations. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 145–149, Melbourne, Australia, 2014e. URL `http://www.aclweb.org/anthology/U14-1021`.

Shervin Malmasi and Mark Dras. From Visualisation to Hypothesis Construction for Second Language Acquisition. In *Proceedings of TextGraphs-9: the Workshop on Graph-based Methods for Natural Language Processing*, pages 56–64, Doha, Qatar, October 2014f. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W14-3708`.

Shervin Malmasi and Mark Dras. Large-scale Native Language Identification with Cross-Corpus Evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2015)*, pages 1403–1409, Denver, CO, USA, June 2015a. Association for Computational Linguistics. URL `http://aclweb.org/anthology/N15-1160`.

Shervin Malmasi and Mark Dras. Automatic Language Identification for Persian and Dari texts. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, pages 59–64, Bali, Indonesia, May 2015b. 16

Shervin Malmasi and Mark Dras. Language Identification using Classifier Ensembles. In *Proceedings of LT4VarDial - Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, Hissar, Bulgaria, 9 2015c.

Shervin Malmasi and Mark Dras. Multilingual Native Language Identification. *Natural Language Engineering*, FirstView:1–53, December 2015d. ISSN 1469-8110. doi: 10.1017/S1351324915000406.

Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. NLI Shared Task 2013: MQ Submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1716`. 47, 145

Shervin Malmasi, Mark Dras, and Irina Temnikova. Norwegian Native Language Identification. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2015)*, pages 404–412, Hissar, Bulgaria, September 2015a. URL `http://www.aclweb.org/anthology/R15-1053`.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, pages 209–217, Bali, Indonesia, May 2015b. 16

Shervin Malmasi, Joel Tetreault, and Mark Dras. Oracle and Human Baselines for Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–178, Denver, Colorado, June 2015c. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-0620`.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Evaluation in information retrieval. In *Introduction to Information Retrieval*, pages 151–175. Cambridge university press Cambridge, 2008. 83

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`. 113

Fethi Mansouri. Agreement morphology in Arabic as a second language. *Cross-linguistic aspects of Processability Theory*, pages 117–253, 2005. 110

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. 125

Melvin Earl Maron. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, 8 (3):404–417, 1961. 14

Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97, 2013. 125

Gerald R McMenamin. *Forensic linguistics: Advances in Forensic Stylistics*. CRC press, 2002. 6, 7

Paul Meara. The bilingual lexicon and the teaching of vocabulary. *The bilingual lexicon*, pages 279–297, 1993. 52, 53

John CP Milton and Elza Shuk-ching Tsang. A corpus-based study of logical connectors in EFL students' writing: directions for future research. In *Studies in lexis. Proceedings of a seminar on lexis organized by the Language Centre of the HKUST, Hong Kong (Language Centre, HKUST, Hong Kong, 1993)*, 1993. 4

Ministry of Labour. Hallituksen maahanmuuttopoliittinen ohjelma. *Tyhallinnon julkaisu 371*, 2006. 102

Tomoya Mizumoto, Yuta Hayashibe, Keisuke Sakaguchi, Mamoru Komachi, and Yuji Matsumoto. NAIST at the NLI 2013 Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 134–139, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1717`. 43

Will Monroe, Spence Green, and Christopher D Manning. Word segmentation of informal Arabic with domain adaptation. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, 2014. 114

Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48, 1991. 162

Mary Morzinski. *The Linguistic influence of Polish on Joseph Conrad's style*. Columbia University Press, New York, NY, 1994. 160

Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA, US, 1964. 19, 31

Shirin Murphy. Second language transfer during third language acquisition. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 3(1), 2005. 131, 187

Steven Myers. Introduction to phishing. In Markus Jakobsson and Steven Myers, editors, *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*, chapter 1, pages 1–29. John Wiley & Sons, Inc., 2007. 3

Carol Myers-Scotton. *Multiple voices: An Introduction to Bilingualism*. Wiley-Blackwell, 2005. 48, 49

Nadja Nesselhauf. Learner corpora and their potential for language teaching. *How to use corpora in language teaching*, page 125, 2004. 104

Garrett Nicolai and Grzegorz Kondrak. Does the phonology of L1 show up in L2 texts? In *Proceedings of The 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 854–859, 2014. 35, 48, 78

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Lei Yao, and Grzegorz Kondrak. Cognate and Misspelling Features for Natural Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 140–145, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1718`. 43, 47

Helle Lykke Nielsen. On acquisition order of agreement procedures in Arabic learner language. *Al-Arabiyya*, 30:49–93, 1997. 110

Tanja Nieminen. Becoming a new Finn through language: non-native English-speaking immigrants' views on integrating into Finnish society, 2009. 102

John M. Norris. Statistical Significance Testing in Second Language Research: Basic Problems and Suggestions for Reform. *Language Learning*, 65(S1):97–126, 2015. ISSN 1467-9922. 134

Joel Nothman, Tara Murphy, and James R Curran. Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 612–620. Association for Computational Linguistics, 2009. 92

Terence Odlin. *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge University Press, Cambridge, UK, 1989. 51, 79, 97, 141

Lourdes Ortega. *Understanding Second Language Acquisition*. Hodder Education, Oxford, UK, 2009. 2, 66, 136, 137

Patric Östergård. A New Algorithm for the Maximum-Weight Clique Problem. *Electronic Notes in Discrete Mathematics*, 3:153–156, May 1999. 152

Agnieszka Otwinowska-Kasztelanic. Raising awareness of cognate vocabulary as a strategy in teaching english to polish adults. *International Journal of Innovation in Language Learning and Teaching*, 3(2):131–147, 2009. 53

Nikunj C Oza and Kagan Tumer. Classifier ensembles: Select real-world applications. *Information Fusion*, 9(1):4–20, 2008. 57

Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey, may 2012. ISBN 978-2-9517408-7-7. 114

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008. 16

Betsy Parrish. A New Look at Methodologies in the Study of Article Acquisition for Learners of ESL. *Language Learning*, 37(3):361–384, 1987. 153

Diane Pecorari. Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing*, 12(4):317–345, 2003. 168

Ria Perkins. *Linguistic identifiers of L1 Persian speakers writing in English: NLID for authorship analysis*. PhD thesis, Aston University, 2014. 6, 7, 47

John G Peters. *Joseph Conrad's Critical Reception*. Cambridge University Press, 2013. 161

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. 125

Lev Pevzner and Marti A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist.*, 28(1):19–36, March 2002. ISSN 0891-2017. doi: 10.1162/089120102317341756. URL `http://dx.doi.org/10.1162/089120102317341756`. 164

Ari Pirkola. Morphological typology of languages for IR. *Journal of Documentation*, 57(3):330–348, 2001. 110

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA, 2000. 60, 73

Robi Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6 (3):21–45, 2006. 58, 59, 83

Marius Popescu and Radu Tudor Ionescu. The Story of the Characters, the DNA and the Native Language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 270–278, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1735`. 45, 46, 48

Matt Post and Daniel Gildea. Bayesian learning of a Tree Substitution Grammar. In *Proceedings of the ACL-IJCNLP 2009 Conference*, pages 45–48, Suntec, Singapore, 2009. Association for Computational Linguistics. 61

Matt Post and Daniel Gildea. Bayesian tree substitution grammars as a usage-based approach. *Language and speech*, 56(3):291–308, 2013. 33

Delip Rao, Michael Paul, Clayton Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. Hierarchical Bayesian models for latent attribute detection in social media. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM, pages 598–601, July 2011. 20

Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 1996)*, volume 1, pages 133–142, 1996. 32

Jack C. Richards. A non-contrastive approach to error analysis. *ELT Journal*, 25(3):204–219, 1971. 49, 50

Jack C. Richards and Theodore S. Rodgers. *Approaches and methods in language teaching*. Cambridge University Press, 2014. 137

Matthew Richardson, Amit Prakash, and Eric Brill. Beyond PageRank: machine learning for static ranking. In *Proceedings of the 15th international conference on World Wide Web*, pages 707–715. ACM, 2006. 73

Hakan Ringbom. Lexical transfer in L3 production. Cenoz and Jessner (2001), pages 59–68. 131, 187

Daniel Robertson. Variability in the use of the English article system by Chinese learners of English. *Second Language Research*, 16(2):135–172, 2000. 143

Monica Rogati and Yiming Yang. High-performing feature selection for text classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 659–661. ACM, 2002. 140

Ute Römer. *Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics*, volume 18. John Benjamins Publishing, 2005. 54

Heath Rose and Lorna Carson. Introduction. *Language Learning in Higher Education*, 4(2):257–269, 2014. 102

Alla Rozovskaya and Dan Roth. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 924–933, Portland, Oregon, USA, June 2011. 6

Karin C. Ryding. Teaching Arabic in the United States. In Kassem M Wahba, Zeinab A Taha, and Liz England, editors, *Handbook for Arabic language teaching professionals in the 21st century*. Routledge, 2013. 102

Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In *Machine learning and knowledge discovery in databases*, pages 313–325. Springer, 2008. 139

Geoffrey Sampson. The SUSANNE corpus. *ICAME Journal*, 17(125127):116, 1993. 125

Mark Sandler. On the use of linear programming for unsupervised text classification. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 256–264. ACM, 2005. 14

Beatrice Santorini. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). Technical report, 1990. 32

Jyoti Sanyal. *Indlish: The Book for Every English-Speaking Indian*. Viva Books Private Limited, 2007. 141

Jacquelyn Schachter. An error in error analysis. *Language Learning*, 27:205–214, 1974. 51

Anne Schiller, Simone Teufel, and Christine Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. *Manuscript, Universities of Stuttgart and Tübingen*, 1995. 114

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK, 1994. 32

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. 124

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002. 14, 17

Mark D Shermis and Jill C Burstein. *Automated essay scoring: A cross-disciplinary perspective*. Psychology Press, 2003. 17

Ling Shi. Cultural backgrounds and textual appropriation. *Language Awareness*, 15(4):264–282, 2006. 168

Peter Siemen, Anke Lüdeling, and Frank Henrik Müller. Falko-ein fehlerannotiertes lernerkorpus des deutschen. *Proceedings of Konvens 2006*, 2006. 107

Kirsti Siitonen. Learners' dilemma: an example of complexity in academic Finnish. The frequency and use of the E infinitive passive in L2 and L1 Finnish. *AFinLA-e: Soveltavan kielitieteen tutkimuksia*, (6):134–148, 2014. 102

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D08-1027`. 86

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009. 168

M. Kendra Sun-Alperin and Min Wang. Spanish-speaking children's spelling errors with English vowel sounds that are represented by different graphemes in English and Spanish words. *Contemporary Educational Psychology*, 33(4):932–948, 2008. 50

Ben Swanson. Exploring Syntactic Representations for Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 146–151, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1719`. 44

Ben Swanson and Eugene Charniak. Extracting the Native Language Signal for Second Language Acquisition. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–94, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N13-1009`. 47, 137, 151

Ben Swanson and Eugene Charniak. Data Driven Language Transfer Hypotheses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-14)*, page 169, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. 22, 137

Benjamin Swanson and Eugene Charniak. Native Language Detection with Tree Substitution Grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 193–197, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P12-2038`. 32, 33, 38, 56, 136

Henry Sweet. *The practical study of languages: A guide for teachers and learners*. Oxford University Press, London, 1899. 51

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12, 2013. 125

Joel Tetreault and Martin Chodorow. Towards Automatically Acquiring Models of ESL Errors. In *CALICO Workshop on Automatic Analysis of Learner Language (AALL '09)*, 2009. 10

Joel Tetreault, Jennifer Foster, and Martin Chodorow. Using Parse Features for Preposition Selection and Error Detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 353–358, 2010. 10

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of COLING 2012*, pages 2585–2602, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL `http://www.aclweb.org/anthology/C12-1158`. 33, 34, 38, 56, 60, 61, 66, 72, 73, 77, 93, 112, 118, 123, 128, 132

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 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, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1706`. 22

Kai Ming Ting and Ian H. Witten. Issues in Stacked Generalization. *Journal of Artificial Intelligence Research*, 10:271–289, 1999. 187

Teresa Tinsley. Languages: the state of the nation: demand and supply of language skills in the UK. Technical report, 2013. 137

Elena Tognini-Bonelli. *Corpus linguistics at work*, volume 6. John Benjamins, 2001. 54

Laura Mayfield Tomokiyo and Rosie Jones. You're not from round here, are you? Naive Bayes detection of non-native utterance text. In *Proceedings of the Second North American Chapter of the Association for Computational Linguistics*, NAACL '01, pages 239–246, 2001. 35

Rosemary Torney, Peter Vamplew, and John Yearwood. Using psycholinguistic features for profiling first language of authors. *Journal of the American Society for Information Science and Technology*, 63(6):1256–1269, 2012. 38, 118

Linda Tsung and Ken Cruickshank. *Teaching and Learning Chinese in Global Contexts: CFL Worldwide*. Bloomsbury Publishing, 2011. 102

Oren Tsur and Ari Rappoport. 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, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W07/W07-0602`. 35, 48, 64, 78, 141

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. Identifying the L1 of non-native writers: the CMU-Haifa system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 279–287, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1736`. 46

Joost Van Doremalen, Catia Cucchiarini, and Helmer Strik. Optimizing automatic speech recognition for low-proficient non-native speakers. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010:2, 2010. 188

Bertus Van Rooy and Lande Schäfer. The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies*, 20(4):325–335, 2002. 72

Dimitra Vergyri, Katrin Kirchhoff, Kevin Duh, and Andreas Stolcke. Morphology-based language modeling for Arabic speech recognition. In *INTERSPEECH*, volume 4, pages 2245–2248, 2004. 110

Nigel Vincent. Italian. In Bernard Comrie, editor, *The world's major languages*, pages 233–252. Routledge, 2009. 105

Kassem M Wahba, Zeinab A Taha, and Liz England. *Handbook for Arabic language teaching professionals in the 21st century*. Routledge, 2013. 102

Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26 th International Conference on Machine Learning*, pages 1105–1112, Montreal, Canada, 2009. 175

Chong Wang and David M Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in Neural Information Processing Systems*, pages 1982–1989, 2009. 179

Maolin Wang, Shervin Malmasi, and Mingxuan Huang. The Jinan Chinese Learner Corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 118–123, Denver, Colorado, June 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W15-0614`. 108

Matthijs J Warrens. On association coefficients for 2× 2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, 73(4):777–789, 2008. 73

Ben Wellner, James Pustejovsky, Catherine Havasi, Anna Rumshisky, and Roser Sauri. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 117–125. Association for Computational Linguistics, 2009. 73

Sarah Williams and Bjorn Hammarberg. Language switches in L3 production: Implications for a polyglot speaking model. *Applied linguistics*, 19(3):295–333, 1998. 131

Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang. Introduction to Chinese Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 2(1):1–148, 2009. 102

Sze-Meng Jojo Wong. *Native Language Identification Incorporating Syntactic Knowledge*. PhD thesis, Macquarie University, Sydney, Australia, 2012. 3

Sze-Meng Jojo Wong and Mark Dras. Contrastive Analysis and Native Language Identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia, December 2009a. URL `http://www.aclweb.org/anthology/U09-1008`. 31, 36, 37

Sze-Meng Jojo Wong and Mark Dras. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia, December 2009b. 118

Sze-Meng Jojo Wong and Mark Dras. Exploiting Parse Structures for Native Language Identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D11-1148`. 32, 34, 36, 38, 56, 62, 136

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. Topic modeling for native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 115–124, Canberra, Australia, December 2011. 36

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. Exploring Adaptor Grammars for Native Language Identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 699–709, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D12-1064`. 34, 38, 61, 136, 170

Michał Woźniak, Manuel Graña, and Emilio Corchado. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16:3–17, 2014. 57

Ching-Yi Wu, Po-Hsiang Lai, Yang Liu, and Vincent Ng. Simple Yet Powerful Native Language Identification on TOEFL11. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 152–156, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-1720`. 43, 115

Fei Xia. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). Technical report, 2000. 114

Suying Yang and Yue Yuan Huang. The impact of the absence of grammatical tense in L1 on the acquisition of the tense-aspect system in L2. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 42(1):49–70, 2004. 50

Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, pages 412–420, 1997. 139

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 180–189, 2011. 16, 147

Helen Yannakoudakis, Ted Briscoe, and Theodora Alexopoulou. Automating Second Language Acquisition Research: Integrating Information Visualisation and Machine Learning. In *Proc. EACL Workshop of LINGVIS & UNCLH*, pages 35–43, 2012. 146, 147, 148, 150, 151, 152, 156

Richard Young. Form-Function Relations in Articles in English Interlanguage. In R. Bayley and D. R. Preston, editors, *Second Language Acquisition and Linguistic Variation*, pages 135–175. John Benjamins, Amsterdam, The Netherlands, 1996. 153

George Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, pages 579–652, 1912. 73

Yi Zhang, Samuel Burer, and W Nick Street. Ensemble pruning via semi-definite programming. *The Journal of Machine Learning Research*, 7:1315–1338, 2006. 187

Hongqin Zhao and Jianbin Huang. Chinas policy of Chinese as a foreign language and the use of overseas Confucius Institutes. *Educational Research for Policy and Practice*, 9(2):127–142, 2010. 102

George Kingsley Zipf. *Human Behavior And The Principle Of Least Effort: An Introduction To Human Ecology*. Addison-Wesley Press, 1949. 29

Sven Meyer Zu Eissen and Benno Stein. Intrinsic Plagiarism Detection. In *Advances in Information Retrieval*, pages 565–569. Springer, 2006. 168