Natural Language Engineering **23** (2): 163–215. © Cambridge University Press 2015
doi:10.1017/S1351324915000406

Multilingual native language identification

SHERVIN MALMASI and MARK DRAS

Centre for Language Technology, Department of Computing, Macquarie University, Sydney, Australia
e-mail: shervin.malmasi@mq.edu.au, mark.dras@mq.edu.au

(*Received 11 March 2015; revised 9 October 2015; accepted 12 October 2015; first published online 2 December 2015*)

Abstract

We present the first comprehensive study of Native Language Identification (NLI) applied to text written in languages other than English, using data from six languages. NLI is the task of predicting an author's first language using only their writings in a second language, with applications in Second Language Acquisition and forensic linguistics. Most research to date has focused on English but there is a need to apply NLI to other languages, not only to gauge its applicability but also to aid in teaching research for other emerging languages. With this goal, we identify six typologically very different sources of non-English second language data and conduct six experiments using a set of commonly used features. Our first two experiments evaluate our features and corpora, showing that the features perform well and at similar rates across languages. The third experiment compares non-native and native control data, showing that they can be discerned with 95 per cent accuracy. Our fourth experiment provides a cross-linguistic assessment of how the degree of syntactic data encoded in part-of-speech tags affects their efficiency as classification features, finding that most differences between first language groups lie in the ordering of the most basic word categories. We also tackle two questions that have not previously been addressed for NLI. Other work in NLI has shown that ensembles of classifiers over feature types work well and in our final experiment we use such an oracle classifier to derive an upper limit for classification accuracy with our feature set. We also present an analysis examining feature diversity, aiming to estimate the degree of overlap and complementarity between our chosen features employing an association measure for binary data. Finally, we conclude with a general discussion and outline directions for future work.

1 Introduction

The task of determining an author's native language (L1) based on their writing in a non-native or second language (L2) is known as Native Language Identification (NLI). NLI works by identifying language use patterns that are common to certain groups of speakers that share the same L1. The general framework of an NLI system is depicted in Figure 1. This process is underpinned by the presupposition that an author's linguistic background will dispose them towards particular language production patterns in their subsequently learnt languages, as influenced by their mother tongue. This relates to the issue of *Cross-linguistic Influence* (CLI), and will be discussed in Section 2.1.

Most studies conducted to date approach NLI as a multiclass supervised classification task. In this experimental design, the L1 metadata are used as class labels

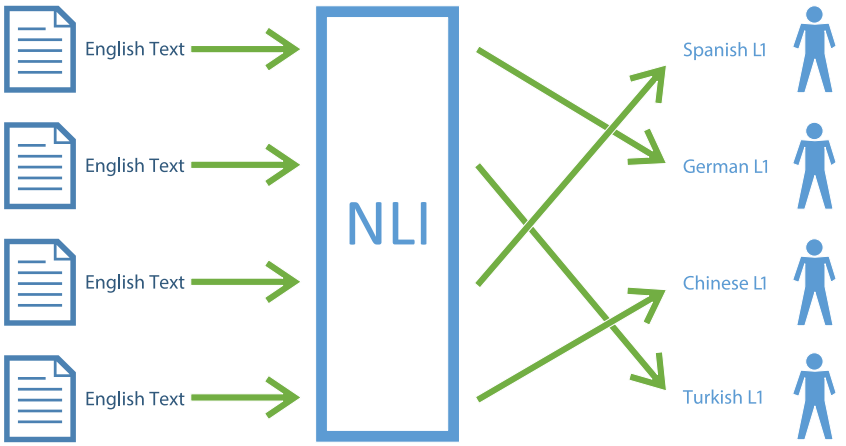

Fig. 1. (Colour online) An example of an NLI system that attempts to classify the native languages (L1) of the authors of non-native (L2) English texts.

and the individual writings are used as training and testing data. Using lexical and syntactic features of increasing sophistication, researchers have obtained good results under this paradigm. Recently, it has also been shown that these machine learning methods outperform human experts for this task (Malmasi *et al.* 2015b).

NLI technology has practical applications in various fields. One potential application of NLI is in the field of forensic linguistics (Gibbons 2003; Coulthard and Johnson 2007), a juncture where the legal system and linguistic stylistics intersect (McMenamin 2002; Gibbons and Prakasam 2004). 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 L1 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. NLI can be applied in such contexts to glean information about the discriminant L1 cues in an anonymous text. One recently studied example is the analysis of extremist related activity on the web (Abbasi and Chen 2005).

Accordingly, we can see that from a forensic point of view, NLI can be a useful tool for intelligence and law enforcement agencies. In fact, recent NLI research such as that related to the work presented by Perkins (2014) has already attracted interest and funding from intelligence agencies (Perkins 2014, 17).

While NLI has such applications in security, most research has a strong linguistic motivation relating to language teaching and learning. Rising numbers of language learners have led to an increasing need for language learning resources, which has in turn fuelled much of the language acquisition research of the past decade.

In connection to the field of Second Language Acquisition (SLA) research, NLI can be used to identify the most challenging aspects of a language for learners from

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

specific backgrounds. In this context, by identifying L1-specific language usage and error patterns, NLI can be used to better understand SLA and develop teaching methods, instructions and learner feedback that is specific to their mother tongue. This tailored evaluation can be derived from language-specific models, whereby learners are provided with customized and specific feedback, determined by their L1. 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). The application of these tools and scientific methods like NLI could potentially assist researchers in creating effective teaching practices and is an area of active research.

In conjunction with SLA, researchers are interested in the nature and degree to which an L1 affects the acquisition and production of other consequently learnt language (Ortega 2009, 31). NLI-based analyses could be used to help researchers in linguistics and cognitive science to better understand the process of L2 acquisition and language transfer effects. Such analyses are often done manually in SLA, and are difficult to perform for large corpora. Smaller studies can yield poor results due to the sample size, leading to extreme variability (Ellis 2008). Recently, researchers have noted that Natural Language Processing (NLP) has the tools to use large amounts of data to automate this analysis using complex feature types, thereby motivating studies in NLI.

1.1 Moving beyond English NLI

While it has attracted significant attention from researchers, almost all of the NLI research to date 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 L1 of over 400 million people in the US, UK 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 L2 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 (Nieminen 2009). This, in turn, has motivated research in studying the acquisition of Finnish to identify the most challenging aspects of the process.[1]

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

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

Another such example is that of Chinese. Interest in learning Chinese is rapidly growing, leading to increased research in teaching Chinese as a foreign language and the development of related resources such as learner corpora (Chen, Wang and Cai 2010). This booming growth in Chinese language learning (Zhao and Huang 2010; Rose and Carson 2014), 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, Taha and England 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 here vary significantly with respect not only to English, but also amongst themselves, in various linguistic subsystems; these differences are detailed in Section 3. In this regard, the current study aims to assess whether these differences significantly impact NLI performance for different L2s.

1.2 Goals and objectives

There are several aims of the present research relating to various 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 type of features that have been common to almost all NLI approaches to date for 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 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 here.

Another issue that arises in this type of multilingual research is the use of multiple part-of-speech tagsets developed for different languages. Differences in the granularity of the tags mean that they are often not directly comparable. This is 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. However, no attempts have been made to measure the degree of dependence, overlap and complementarity between different features. Accordingly, another aim of the present inquiry is to apply and evaluate a suitable method for measuring and quantifying this interfeature diversity and to assess how the results compare across languages.

The final objective of the paper 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. Other work in NLI has shown that ensembles of classifiers work well and in our final experiment we use an oracle classifier to derive an upper limit for classification accuracy with our feature set. This is something that has not been previously investigated and can be a helpful baseline to guide and interpret future research. 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 are reflected across languages, which is of importance to multilingual research in this area.

1.3 Paper outline

The rest of this paper is organized as follows. We begin by reviewing previous research in Section 2. The data and corpora we use are presented in Section 3. Our methodology is outlined in Section 4 and the feature types used therein are described in Section 5. The descriptions and results from our experiments are detailed in Sections 6–11 and followed by a general discussion in Section 12 that summarizes the conclusions of our experiments and outlines some directions for future research.

2 Related work

2.1 Cross-linguistic influence

CLI, also referred to as language transfer, is one of the major topics in the field of SLA. It has been said that being a speaker of some specific L1 can have direct and indirect consequences on an individual's usage of some later-learned language (Jarvis and Crossley 2012), and this is the effect that is studied under the heading of CLI. With this in mind, SLA research aims to find distributional differences in language use 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).

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

We now briefly 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, as described below.

2.1.1 Positive transfer

This type of transfer is generally facilitated by similarities between the native tongue and L2s. 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 (NSs) may find it easier to learn each other's languages in comparison with those that speak a language which utilizes a phonetic alphabet.

2.1.2 Underuse (avoidance)

The underutilization of particular linguistic structures is known as avoidance. While the existence of this phenomenon 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 mostly avoiding the use of such structures. Another reason for avoidance may be the inherent complexity of the structures themselves (Gass and Selinker 2008).

2.1.3 Overuse

The above-mentioned 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) also discusses how 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*.

These are the types of patterns that SLA researchers attempt to uncover using learner corpora (Granger 2009). While there are some attempts in SLA to use computational approaches on small-scale data, e.g. Chen (2013) and Lozanó and Mendikoetxea (2010), these still use fairly elementary computational tools, including mostly manual approaches to annotation.

One such 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, using 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 reannotation of the data.

Among the efflorescence of work in Computational Linguistics, researchers have turned their attention to investigating these phenomena through predictive computational models. The majority of these models are based on the aforementioned theories relating to learner interlanguage. NLI is one such area where work has focused on automatic learner L1 classification using machine learning with large-scale data and sophisticated linguistic features (Tetreault *et al.* 2012). Other work has linked this directly to issues of interest in SLA: linking errors and L1 (Kochmar 2011), methods for proposing potential SLA hypotheses (Swanson and Charniak 2013; Malmasi and Dras 2014b), and so on. This is also the approach pursued in this work, where a large learner corpora of different languages will be used in conjunction with automatic linguistic annotation and machine learning methods.

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 (Ortega 2009; Richards and Rodgers 2014). 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 easy task, but a good starting point is to reduce those idiosyncratic language use patterns caused by the influence of the L1. 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, White and Bell (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 activities which were tested with francophone learners of English at a school in Montreal, Quebec, Canada. The cross-linguistic awareness material was developed by identifying commonalities between

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

French and English. Next, a set of 11 cross-linguistic awareness teaching packages were developed and piloted in an intensive year-long English as a Second Language (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, non-contrastive form-focused instruction and contrastive analysis and translation. The contrastive analysis and translation 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 contrastive analysis and translation 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.

2.3 Native language identification

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 surge of interest, coupled with the inaugural shared task in 2013 (Tetreault *et al.* 2013), has resulted in NLI becoming a well-established NLP task. We point out just the previous research on the task relevant to the present article.

The earliest work on detecting L2 is that of Tomokiyo and Jones (2001) whose main aim was to detect non-native speech using part-of-speech and lexical *n*-grams, and to also determine the L1 of the non-native speakers (NNSs). They were able to achieve 100 per cent accuracy in their study, which included six Chinese and thirty-one Japanese speakers.

Koppel *et al.* (2005a; 2005b) established the text classification paradigm now widely used in the area. Texts ranging from 500 to 850 words from five L1s were selected from the first version of the International Corpus of Learner English (ICLE) (Granger *et al.* 2009). They used a set of syntactic, lexical and stylistic features that included function words, character *n*-grams and part-of-speech (POS) bigrams, together with spelling mistakes. Using an Support Vector Machine (SVM) classifier, the achieved a classification accuracy of 80 per cent with ten-fold cross-validation – a strong result given the 20 per cent chance baseline.

Wong and Dras (2011) proposed exploiting parse structures for NLI. They explored the usefulness of syntactic features in a broader sense by characterizing syntactic errors with cross sections of parse trees obtained from statistical parsing. More specifically, they utilized 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 performed significantly better than lexical features alone, giving the best performance on the ICLE (v.2) dataset at the time. Other syntactic information, in such forms as Tree Substitution Grammars (Swanson and Charniak 2012) or dependency relations (Tetreault *et al.* 2012), have subsequently also been used.

This set of core function word and POS-based features were used by most follow-up studies, including Tsur and Rappoport (2007), Estival *et al.* (2007), Kochmar (2011), Brooke and Hirst (2011, 2012a) and Wong, Dras and Johnson (2012).

Tetreault *et al.* (2012) proposed the use of classifier ensembles for NLI. In their study, they used an ensemble of logistic regression learners each trained on a wide range of features that included POS *n*-grams, function words, spelling errors and writing quality markers, amongst others. This was in contrast with previous work that had combined all features in a single space. The set of features used here was also the largest of any NLI study to date. With this system, the authors reported state of the art accuracies of 90.1 per cent and 80.9 per cent on the ICLE and TOEFL11 corpora (introduced in this work and now standard – see Section 3.7.2), respectively.

We note that this approach, and previous work, made no attempt to measure the diversity between feature types to determine if any feature pairs are capturing the same information. This is an important factor to consider, particularly when building ensembles with many feature types.

Increased interest in NLI brought unprecedented levels of research focus and momentum, resulting in the first NLI shared task being held in 2013.[2] The shared task aimed to facilitate the comparison of results by providing a large NLI-specific dataset and evaluation procedure, to enable direct comparison of results achieved through different methods. Overall, the event was considered a success, drawing twenty-nine entrants and experts from not only Computational Linguistics, but also SLA. The best teams achieved accuracies of around 80 per cent on this 11-class classification task where the great majority of entries used standard features such as POS *n*-grams. A detailed summary of the results can be found in Tetreault *et al.* (2013).

We can identify a number of relevant trends from this survey of NLI literature. First, we observe that function words and POS *n*-grams constitute a core set of standard features for this task: these will be our fundamental features as well. Second, facilitated by the large body of learner English data that has accumulated over the last few decades, NLI researchers have focused almost exclusively on English. With this in mind, one of the central contributions of this work is the extension of NLI to additional, non-English L2s such as Italian and German. Third, there are a number of issues worth closer analysis, including the above-mentioned issue of feature diversity.

[2] Organized by the Educational Testing Service and colocated with the eighth instalment of the Building Educational Applications Workshop at NAACL/HLT 2013. `http://sites.google.com/site/nlisharedtask2013/`

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

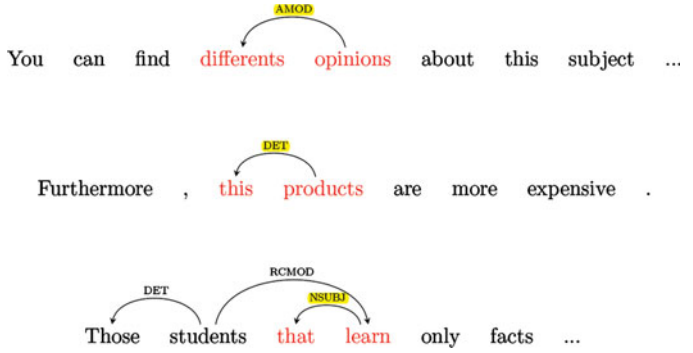

Fig. 2. (Colour online) Three of the most common overuse patterns found in the writing of L1 Spanish learners. They show erroneous pluralization of adjectives, determiner misuse and overuse of the word *that*.

2.3.1 *From NLI to language transfer hypotheses*

NLI methods have also been extended to investigating language transfer and cross-linguistic effects, as described in this section. The learner corpora are used to extract potential language transfer effects between the languages in each corpus using data-driven methodology such as the one proposed by Swanson and Charniak (2014). Malmasi and Dras (2014b) also propose such a method using SVM weights and apply it to generate potential language transfer hypotheses from the writings of English learners in the TOEFL11 corpus.

For Spanish L1 authors, they extract both underuse and overuse lists of syntactic dependencies. The top three overuse rules show the word *that* is very often used as the subject of verbs. 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*. Another 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*'. Yet another rule shows that the adjectival modifier of a plural noun is being incorrectly pluralized to match the noun in number as would be required in Spanish, for example, 'differents subjects'. Some examples of these dependencies are shown in Figure 2. Turning to the underused features in Spanish L1 texts, they show that four related features rank highly, demonstrating that *these* is not commonly used as a determiner for plural nouns and *which* is rarely used as a subject.

3 Learner corpora and languages

In this section, we outline the data used in this study. 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 to English, the most commonly investigated language in NLI.

Learner corpora – datasets comprised 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).

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

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 NLI (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. For L2 English, the two main datasets are the 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 (Abuhakema *et al.* 2008; Lozano and Mendikoetxea 2013). 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. An ideal NLI corpus should

- have multiple and diverse L1 groups represented,
- be balanced by topic so as to avoid topic bias,[4]
- be balanced in proficiency across the groups,
- contain similar numbers of texts per L1, i.e. be balanced by class,
- be sufficiently large in size to reliably identify intergroup differences.

One key contribution of this work 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 1. Additionally, a listing of the L1 groups and text counts for each corpus can be found in Table 2, which also shows the average text length for documents in each class in tokens, except for Chinese, which is measured in characters.

3.1 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.

[3] TOEFL11, described later in this section, contains c. 4 million tokens in 12,100 texts.
[4] See the end of Section 5 for more details.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

Table 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*

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	314 (176)	206	N
Arabic	ALC	7	155 (76)	329	N
Italian	VALICO	14	210 (105)	2,531	N
German	FALKO	8	404 (135)	221	N
Chinese	JCLC	11	610 (26)	3,216	N
Finnish	LAS2	9	575 (304)	204	N

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 utilize 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 twenty L1 groups are represented in the data, many do not have sufficient data for our purposes. We have selected the top fourteen native languages by the number of available texts as the rest of the classes contain too few texts; these are shown in Table 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.

3.2 German

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

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

Table 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*

	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	
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	
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	

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

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

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.

The largest publicly available selection of German learner texts can be found in the FALKO (***f**ehler**a**nnotierten **L**erner**ko**rpus*) corpus[5] by Siemen *et al.* (2006) and this is the source of the German data used in this work.

It has several subcorpora, including the essay subcorpus (argumentative essays written by learners) and summary subcorpus (text summaries written by learners). It also contains baseline corpora with texts written by German NSs. For the purposes of our experiments, we combine the essay and summary texts, but do not use the longitudinal subcorpus texts. A listing of the L1 groups and text counts of the corpus subset we use can be found in Table 2.

3.3 Spanish

As the Romance language with the greatest number of speakers, Spanish is the official language of some twenty 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, 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[6] comprised of Spanish texts that were produced by students either as course work 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 (European Region Action Scheme for the Mobility of University Students) students from France, Germany and Italy. These students have varied nationalities and backgrounds (56 per cent do not have English as L1).

Each text includes metadata with the following information: the task set, the conditions (exam/course work), the text type (narrative, description, etc.), proficiency

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

[6] 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.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

level (beginner, intermediate or advanced), course book (where known), student identity number, L1 and gender.

A total of twenty 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 2.

3.4 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.

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: 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 the recent development of the Jinan Chinese Learner Corpus (JCLC) (Wang, Malmasi and Huang 2015), the first large-scale corpus of L2 Chinese consisting of university student essays. Learners from fifty-nine

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

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 geographical proximity and links to China.

For this work, we extracted 3.75 million tokens of text from the JCLC in the form of individual sentences.[7] Following the methodology 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.[8]

Although there are over fifty L1s available in the corpus, we choose the top eleven languages, shown in Table 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 eleven 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.

3.5 Arabic

Arabic, part of the Semitic family of languages, is the official language of over twenty countries. It is comprised of many regional dialects with the Modern Standard Arabic 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 leads to 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 (Vergyri *et al.* 2004; Diab 2009). 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.

[7] Full texts are not made available, only individual sentences with the relevant metadata (proficiency/nationality).

[8] Pending permission from the CLC corpus authors, we will attempt to release this Chinese NLI dataset publicly.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

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 interphrasal 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[9] (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, sixty-six 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.[10]

We use the more recent second version of the ALC (Alfaifi, Atwell and Hedaya 2014) as the data for our experiments. While there are sixty-six different L1s in the corpus, the majority of these have fewer than ten texts and cannot reliably be used for NLI. Instead, we use a subset of the corpus consisting of the top seven L1s 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 2.

3.6 Finnish

The final language included in the present work 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 characterizing 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 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

[9] http://www.arabiclearnercorpus.com/
[10] http://www.alcsearch.com/

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

being collected as part of an ongoing project at the University of Turku[11] 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 fifteen different L1 backgrounds. The included L1 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 fifteen different L1s represented in the corpus, the majority of these have fewer than ten texts and cannot reliably be used for NLI. Instead, we use a subset of the corpus consisting of the top seven L1s by number of texts. The languages and document counts in each class are shown in Table 2.

3.7 Other corpora

In this section, we describe a number of other corpora that we use in this work, namely for Experiment 3.

3.7.1 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 NS 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.

3.7.2 The TOEFL11 *corpus*

Initially released as part of the 2013 NLI Shared task, the TOEFL11 corpus (Blanchard *et al.* 2013) is the first dataset designed specifically for the task of NLI and developed with the aim of addressing the 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.

It consists of 12,100 learner texts from speakers of eleven different languages, making it the largest publicly available corpus of non-native English writing. The texts are independent task essays written in response to eight different prompts,[12]

[11] `http://www.utu.fi/fi/yksikot/hum/yksikot/suomi-sgr/tutkimus/tutkimushankkeet/las2/Sivut/home.aspx`

[12] An essay prompt is a statement that sets the topic of the essay, e.g. 'A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught. Use specific reasons and examples to support your answer.'

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

and were collected in the process of administering the Test of English as a Foreign Language (TOEFL®) between 2006 and 2007. The eleven L1s are Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu and Turkish. This dataset was designed specifically for NLI and the authors balanced the texts by topic and L1. Furthermore, the proficiency level of the author of each text (low, medium or high) is also provided as metadata.

Furthermore, as all of the texts were collected through the Education Testing Service's electronic test delivery system, this ensures that all of the data files are encoded and stored in a consistent manner. The corpus was released through the Linguistic Data Consortium in 2013.

3.7.3 The LOCNESS corpus

The Louvain Corpus of Native English Essays (LOCNESS)[13] – 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 NS data in the TOEFL11 corpus.

3.8 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.

3.8.1 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.

3.8.2 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. We also note that some languages may be represented by various character encoding standards,[14] 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.

[13] http://www.learnercorpusassociation.org/resources/corpora/locness-corpus/
[14] For example, GB18030, GBK and GB2312 for Chinese text.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

3.8.3 Unicode normalization

This is the process of converting Unicode strings so that all canonical-equivalent strings[15] have the exact same binary representation.[16] 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 L1s. 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.

3.8.4 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 original 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.

3.8.5 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.

4 Experimental methodology

We also follow the supervised classification approach described in Section 2. We devise and run experiments using several models that capture different types of

[15] 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.

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

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

linguistic information. For each model, features are extracted from the texts and a classifier is trained to predict the L1 labels using the features.

4.1 Classification

We use a linear Support Vector Machine to perform multiclass classification in our experiments. In particular, we use the LIBLINEAR[17] SVM package (Fan *et al.* 2008) which has been shown to be efficient for text classification problems with large numbers of features and documents such as the present work. It has also been demonstrated to be the most effective classifier for NLI in the 2013 NLI Shared Task (Tetreault *et al.* 2013). More specifically, we make use of the L2-regularized L2-loss support vector classification (dual) solver.

4.2 Evaluation

Consistent with most 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 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 $415/3216 = 12.9$ per cent.

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

4.3 NLP tools

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

4.3.1 Chinese and German

For processing these two languages, the Stanford CoreNLP[18] suite of NLP tools (Manning *et al.* 2014) and the provided models were used to tokenize POS tag and parse the unsegmented corpus texts.

4.3.2 Arabic

The tokenization and word segmentation of Arabic is an important preprocessing step for addressing the orthographic issues discussed in Section 3.5. For this task,

[17] http://www.csie.ntu.edu.tw/~cjlin/liblinear/
[18] http://nlp.stanford.edu/software/corenlp.shtml

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

we utilize the Stanford Word Segmenter (Monroe *et al.* 2014).[19] The Arabic texts were POS tagged and parsed using the Stanford Arabic Parser.[20]

4.3.3 Spanish and Italian

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

4.3.4 Finnish

We did not use any NLP tools for Finnish as the corpus we use is already annotated.

5 Features

5.1 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 morphosyntactic information. The assignment of such categories to words in a text adds a level of linguistic abstraction.

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 present them in our work.

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[21] 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 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 Section 9.

[19] http://nlp.stanford.edu/software/segmenter.shtml
[20] http://nlp.stanford.edu/projects/arabic.shtml
[21] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

Table 3. *A listing of the tagsets used for the languages in our experiments, including the size of the tagset*

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 4. *Function word counts for the various languages in our study*

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

5.2 Function words

In contrast to content words, function words do not have any meaning 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[22] 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 fifty to several hundred words, depending on the granularity of the list and also the language in question. Table 4 lists the number of such words that we use for each language.

The English word list was obtained from the Onix Text Retrieval Toolkit.[23] For Chinese, we compiled a list of 449 function words using Chinese language teaching

[22] For example, the largest list used by Wong and Dras (2009) 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.

[23] `http://www.lextek.com/manuals/onix/stopwords1.html`

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

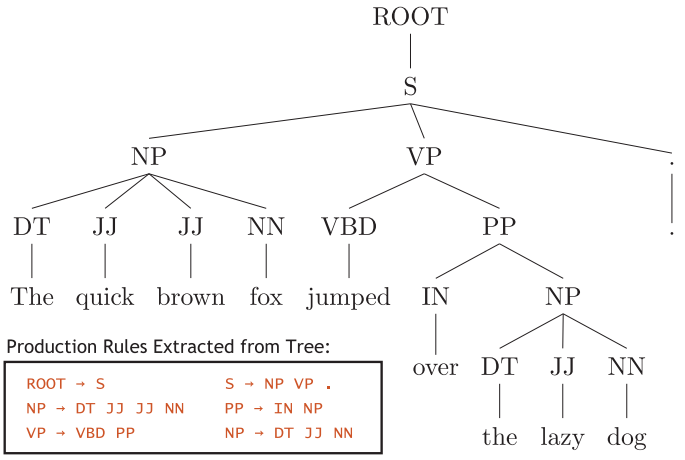

Fig. 3. (Colour online) A constituent parse tree for an example sentence along with the context-free grammar production rules which can be extracted from it.

resources. The complete list can be accessed online.[24] 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.[25]

As seen in Table 4, 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.

5.3 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. The rules are extracted by first generating constituent parses for all sentences. The production rules, excluding lexicalizations, are then extracted. Figure 3 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) and we utilize them as classification features in some of our experiments. 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.

[24] http://comp.mq.edu.au/%7Emadras/research/data/chinese-fw.txt
[25] http://members.unine.ch/jacques.savoy/clef/index.html

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

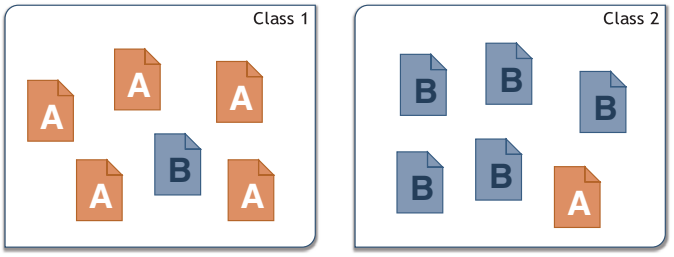

Fig. 4. (Colour online) An example of a dataset that is not balanced by topic: class 1 contains mostly documents from topic A while class 2 is dominated by texts from topic B. Here, a learning algorithm may distinguish the classes through other confounding variables related to topic.

5.4 Unused lexical features

A number of other lexical features that directly use the tokens in a text, including character and word *n*-grams, have also been investigated for NLI. However, the use of these lexical features cannot be justified in all circumstances due to issues with topic bias (Brooke and Hirst 2012a), as we describe here. Topic bias 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 4.

Other researchers like Brooke and Hirst (2012b), however, argue that lexical features cannot be simply ignored. Given the relatively small size of our data and the inability to reach definitive conclusions regarding this, we do not attempt to explore this issue in the present work.

6 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 Section 3.

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). This is an issue that we explore 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

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

features into a single vector to create combined classifiers to assess if a union of the features can yield higher accuracy.

6.1 Results and discussion

The results for all of our languages are included in Table 5. 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).

6.2 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, 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. The combination of both feature representations has also been reported to be effective in previous research (Malmasi *et al.* 2013).[26]

Below, we note language-specific details and analyses.

6.3 Chinese

The results show that POS tags are very useful features here. The trigram frequencies give the best accuracy of 55.60 per cent, 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 Section 3.4. Function word frequency features provide an accuracy of 51.91 per cent, 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 per cent. Again, as for English L2 data, the syntactic substructures

[26] We did not investigate this as it relates to building classifier ensembles, something which is not the focus of this study.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

Table 5. *NLI classification accuracy (per cent) for Chinese (eleven classes), Arabic (seven classes), Italian (fourteen classes), Finnish (nine classes), German (eight classes) and Spanish (six classes). Results are reported using both binary and frequency-based feature representations. The production rules features were only tested on some languages*

	Chinese		Arabic		Italian		Finnish		German		Spanish	
Feature	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

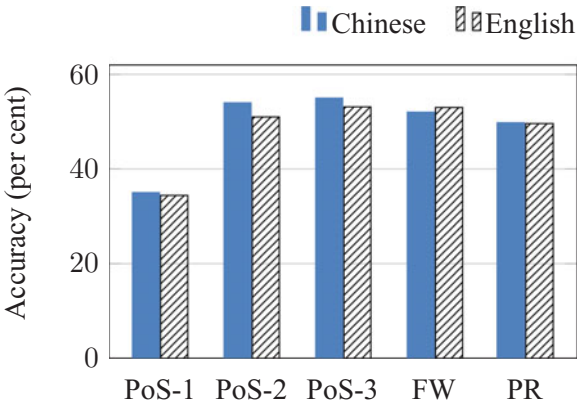

Fig. 5. (Colour online) Comparing feature performance on the CLC and TOEFL11 corpora. POS-1/2/3: POS uni/bi/trigrams, FW: Function words, PR: Production rules.

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 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.

6.4 Arabic

The frequency distributions of the production rules yield 31.7 per cent accuracy and function words achieve 29.2 per cent. While all the models provide results above the baseline, POS tag *n*-grams are the most useful features. Combining all of the models into a single feature space provides the highest accuracy of 41 per cent.

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 per cent, 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

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

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 (2009) which used some similar features on a 7-class English dataset. Despite their use of a much larger dataset,[27] our individual models are only around 10 per cent 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.[28] Based on this, we expect that given similarly sized training data, an Arabic NLI system can achieve similar accuracies.

6.5 Italian

We make use of all fourteen 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.[29] Function words scored 50.1 per cent accuracy and POS trigrams yielded 56.89 per cent. Combining all of the features together improves this to 69.09 per cent, four times higher than the baseline.

6.6 Finnish

Here, we observe that the distribution of function words yields 54.6 per cent 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 per cent 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 per cent, compared to the strongly configurational Chinese and Italian, where the gap is 4–6 per cent.

The combined model provides the highest accuracy of 58.86 per cent, around 4 per cent better than the best single feature type. An interesting difference is that POS unigrams achieve a much lower accuracy of 36.3 per cent.

6.7 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.

[27] Wong and Dras (2009) had 110 texts per class, with average text lengths of more than 600 words.
[28] Equivalent to 1000 texts per L1 class.
[29] Previously, Torney, Vamplew and Yearwood (2012) used sixteen classes from the ICLE.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

6.8 Spanish

The pattern here is most similar to Chinese and Italian.

6.9 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 per cent through to 90 per cent. 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 per cent of it to give a meaningful result, we analyse the curves only for the English TOEFL11 data and our two biggest non-English datasets: Chinese and Italian. The curves are presented in Figure 6.

The curves demonstrate similar rates of learning across languages. We also note that while the relationship between function word and POS trigram features is not 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.

7 Experiment II – comparing languages

The focus of our second experiment is to 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 6 shows the six languages that were selected from each of the three corpora. The number of documents within each class are 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 favoured 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.

7.1 Results

The results for all three languages are shown in Table 7. Each language has a majority class baseline of 16.67 per cent 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.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

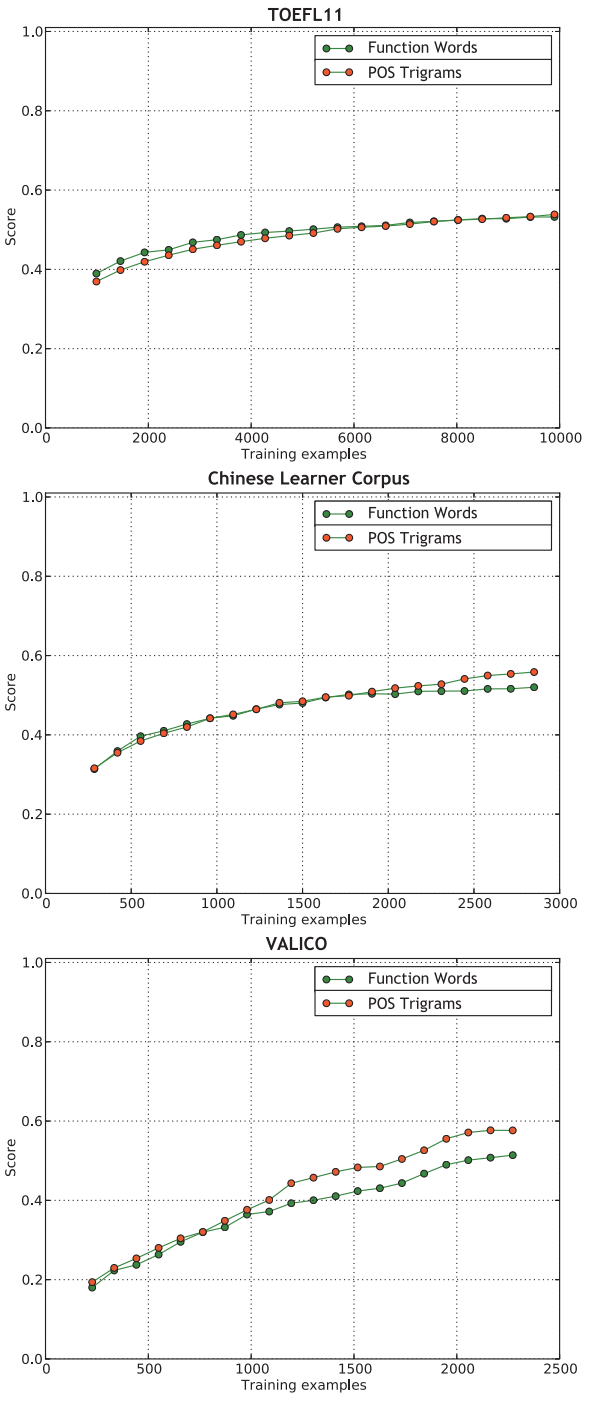

Fig. 6. (Colour online) The learning curves (classification accuracy score versus training set size) for two feature types, Function words and POS trigrams, across three languages: English (Toefl11, row 1), Chinese (CLC, row 2) and Italian (VALICO, row 3).

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

Table 6. *The six L1 classes used for each language in Experiment II*

Language	L1 Classes
Chinese 330 texts per class	Filipino, Indonesian, Thai Laotian, Burmese, Korean
Italian 200 texts per class	French, Japanese, Spanish English, Polish, German
English 1,100 texts per class	French, Japanese, Spanish Hindi, Turkish, Arabic

Table 7. *Comparing classification results across languages*

	Accuracy		
Feature	Chinese	Italian	English
Random baseline	16·67 per cent	16·67 per cent	16·67 per cent
(1) Function words	62·12 per cent	59·24 per cent	63·82 per cent
(2) POS unigrams	47·78 per cent	51·15 per cent	48·59 per cent
(3) POS bigrams	63·14 per cent	63·58 per cent	63·70 per cent
(4) POS trigrams	64·31 per cent	64·66 per cent	65·62 per cent
All features (1–4)	68·14 per cent	67·61 per cent	70·05 per cent

Chinese yields 68.14 per cent accuracy, while Italian and English data obtain 67.61 per cent and 70.05 per cent, respectively. All of these are more than four times higher than the baseline.

7.2 Discussion

These results, shown graphically in Figure 7, 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 Section 6.1, this may be related to the degree of configurationality in these 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.

8 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 NS or NNS author. This should be a feasible task, given the results of the previous experiments. The objective here to

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

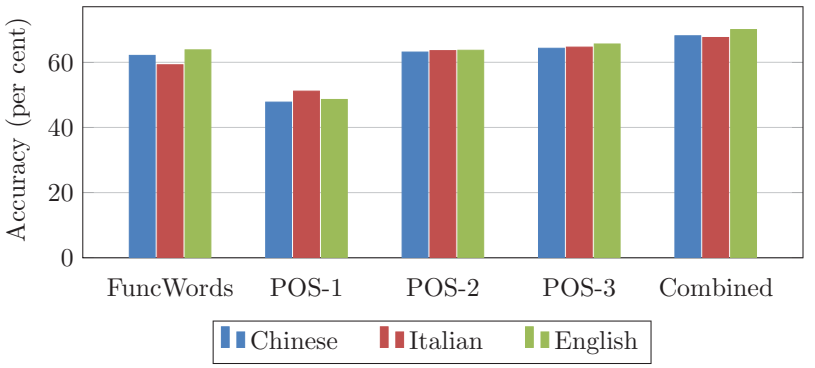

Fig. 7. (Colour online) Performance of our syntactic features (Function words and part-of-speech 1–3 grams) across the three languages.

see to what degree our features can distinguish the writings of NNSs 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: NS author and NNS author. Texts for the NNS class will come from learner corpora of three different languages while data from NS 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[30] from each language[31] listed in Table 2.

For Spanish, we use the CEDEL2 corpus, described in Section 3.7.1. Here, we use 700 NS 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 Section 3.7. This is required as the TOEFL11 corpus does not contain any native control texts. The NS class is composed of 400 NS essays taken from the LOCNESS corpus and the NNS data comes from the TOEFL11 corpus. We sample this data evenly from the eleven L1 non-native classes, selecting thirty-six or thirty-seven 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 per cent. This experiment only uses frequency-based feature value representations and results are reported as classification accuracy under 10-fold cross-validation.

[30] So that the NNS class consists of a similar number of texts from each L1 class.
[31] English only has ten texts, so we include two extra Japanese texts to create a set of hundred documents, with roughly eleven texts from each L1 class.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

Table 8. *Accuracy for classifying texts as native or non-native*

Feature	Accuracy		
	Finnish	Spanish	English
Random baseline	50·00 per cent	50·00 per cent	50·00 per cent
(1) Function words	93·96 per cent	91·12 per cent	94·26 per cent
(2) Part-of-speech unigrams	88·54 per cent	88·71 per cent	87·91 per cent
(3) Part-of-speech bigrams	90·15 per cent	90·35 per cent	91·81 per cent
(4) Part-of-speech trigrams	91·45 per cent	91·35 per cent	92·87 per cent
(5) Production rules	N/A	91·28 per cent	93·61 per cent
All features combined	94·92 per cent	95·23 per cent	96·45 per cent

8.1 Results and discussion

Table 8 shows the results for all three languages, demonstrating that all features greatly surpass the 50 per cent 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 per cent 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 per cent.

For Spanish, all features with the exception of POS unigrams achieve accuracies of over 90 per cent. When combined, the model yields the best accuracy of per cent.

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 8.

We note that although the accuracy increases as amount of data increases, the curve is much flatter than those for NLI in Section 6. However, this is offset by the fact that the curve's starting point is much higher, achieving over 85 per cent accuracy by using only 10 per cent 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 per cent. 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. This was a limitation that we were unable to overcome with the currently available data. In the 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.

One direction for future experiments is the investigation of the relationship between L2 proficiency and the detection accuracy of NNS writing. Previous results by Tetreault *et al.* (2012) show that NLI accuracy decreases as writing proficiency improves and becomes more native-like, but would this pattern hold here as well? Another potential path for future work is to extend this experiment to the

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

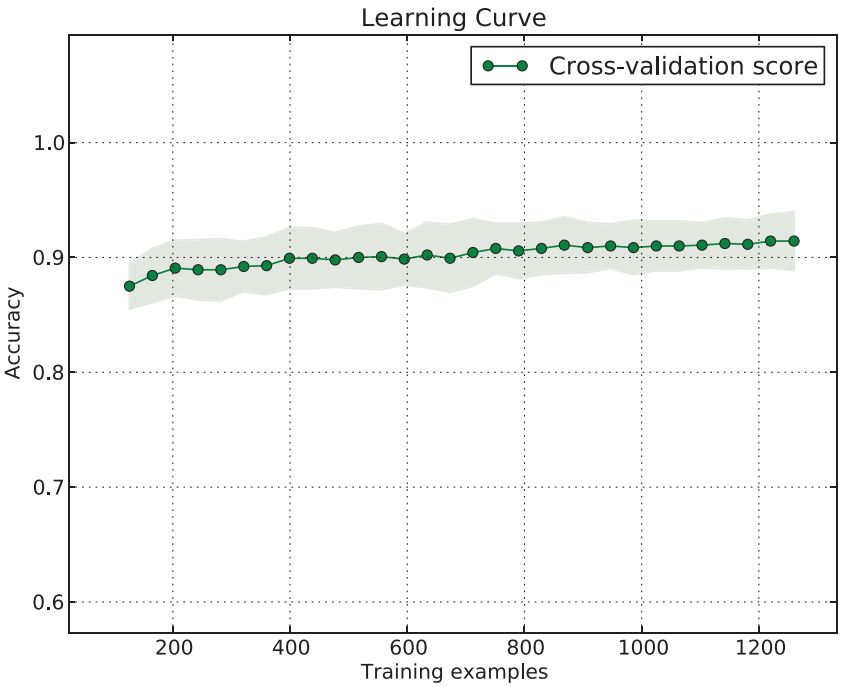

Fig. 8. (Colour online) A learning curve for the Spanish NS author *versus* non-native speaker author classifier trained on POS trigrams. The standard deviation range is also highlighted.

sub-document level to evaluate the applicability of this approach at the paragraph, or even sentence level.

An analysis of what is being learnt by the models here can be useful for understanding the syntactic and stylistic factors that seem to make native writing so easily distinguishable. An exposition of features whose presence, or absence, characterizes non-native writing could also be of use in language teaching and pedagogical research.

It should also be noted that it is 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 be 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.

9 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

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

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 morphosyntactic 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

The present work also makes use of a slew of different tagsets for the different languages, which were outlined in Section 5. 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.

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 also aim to compare the effectiveness of POS tags cross-linguistically. This could also enable us to better understand the results from Section 6 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, Ramirez and Solorio (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.

9.1 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 (UPOS) consisting of twelve coarse POS categories that are considered to be universal across languages.[32]

[32] 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

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

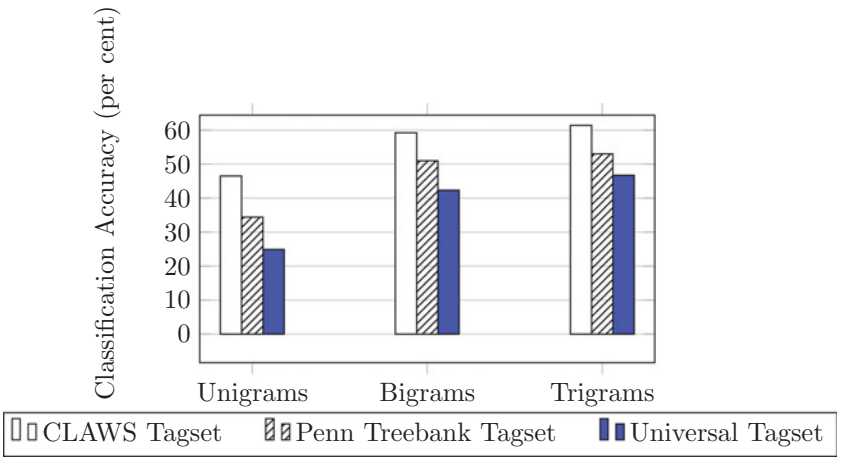

Fig. 9. (Colour online) NLI classification accuracy for L2 English data from the TOEFL11 corpus, using POS *n*-grams extracted with the CLAWS, Penn Treebank and Universal POS tagsets.

We utilize this 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 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 and UPOS.

9.2 Results and discussion

The results for English are shown in Figure 9 and demonstrate that the largest tagset – CLAWS – provides the best classification accuracy. Classification accuracy continues to drop as the tagset gets smaller.

Figures 10 and 11 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.[33]

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 twelve POS tags, the UPOS set retains around 80 per cent 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

for other categories such as abbreviations or foreign words). These categories were derived through analysis of tagsets proposed for twenty-two different languages.

[33] We observe 330 tags in our Italian data while the Penn Chinese Treebank only uses thirty-three tags. The reduction from 330 to 12 tags is steeper, hence the greater drop in accuracy.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

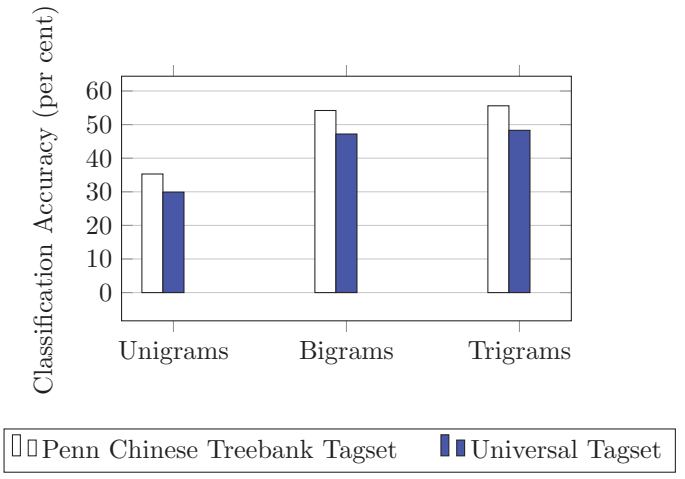

Fig. 10. (Colour online) NLI classification accuracy for the L2 Chinese data, using POS *n*-grams extracted with the Penn Chinese Treebank and Universal POS tagsets.

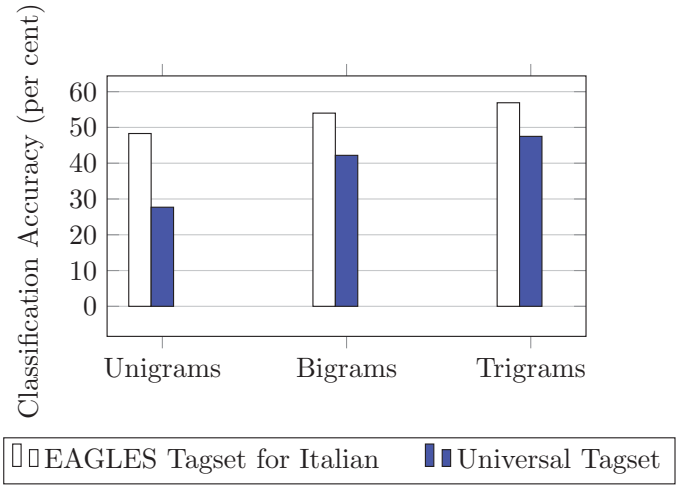

Fig. 11. (Colour online) NLI classification accuracy for the L2 Italian data using POS *n*-grams extracted with the EAGLES and the Universal POS tagsets.

be further investigated by comparing learner data with the same L1 but multiple L2s[34] to find common transfer patterns related to that L1.

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 2014a). 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.

[34] For example, comparing Chinese–English, Chinese–Spanish and Chinese–French.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

Table 9. *Example oracle results for an ensemble of three classifiers*

Instance	True label	Classifier output			Oracle
		C_1	C_2	C_3	
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

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.

10 Experiment V – bounding classification accuracy

Another interesting question about NLI research concerns the maximum potential accuracy that can be obtained for a dataset. More specifically, given a dataset, a selection of features and classifiers, what is an upper bound on the performance that could possibly be achieved by an NLI system that always picks the best candidate?

This question, not previously addressed in the context of NLI to date, is the focus of the next experiment. Such a measure is an interesting and useful upper-limit baseline for researchers to consider when evaluating their work, since obtaining 100 per cent classification accuracy may not be a reasonable or even feasible goal. However, the derivation of such a value is not a straightforward process. In this section, we present an experiment that investigates this issue with the aim of deriving such an upper limit for NLI accuracy.

A possible approach to this question, and one that we employ here, is the use of an 'Oracle' classifier. This method has previously been used to analyse the limits of majority vote classifier combination (Kuncheva *et al.* 2001; Wozniak and Zmyslony 2010). 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 9. 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.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

Table 10. *Oracle classifier accuracy for the three languages in experiment V*

	Italian	Chinese	English
Majority baseline	16·40 per cent	12·90 per cent	09·09 per cent
Our best accuracy	69·09 per cent	70·61 per cent	70·58 per cent
Oracle accuracy	84·35 per cent	87·60 per cent	86·20 per cent

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. Another scenario involves the combination of different learning algorithms,[35] trained on similar features, to form an ensemble in order to evaluate the potential benefits and limits of combining different classification approaches.

In this experiment, we use our feature set on our biggest datasets: Chinese, Italian and English. Following the above-described oracle methodology, we train a single linear SVM classifier for each feature type to create our NLI classifier ensemble, noting that Tetreault *et al.* (2012) found ensembles of classifiers over feature types to produce higher results. We do not experiment with combining different machine learning algorithms 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.

10.1 Results

The oracle results for the three languages are shown in Table 10 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. Analysing 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 per cent of the texts in each dataset. However, even under this best scenario, we should note that not a single classifier is able to correctly predict the label for the remaining per cent 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.

10.2 Discussion

This experiment presented a new type of analysis for predicting the 'potential' upper limit of NLI accuracy on a dataset. This upper limit can vary depending

[35] For example, SVMs, Logistic Regression and String Kernels.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

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.[36]

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.

The method can also be applied in a cross-corpus setting, as described in Section 11.2. Here, the oracle could be useful in measuring the potential cross-corpus accuracy, particularly in settings where the source and target corpora differ significantly in domain, genre or style. The oracle method can help assess how effective the training data are for each target corpus.

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.

We must also bear in mind that these Oracle figures would be produced by an absolutely optimal system that would always make 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. In practice, this type of Oracle measure can be used to guide the process of choosing the pool of classifiers that form an ensemble.

11 Measuring and analysing feature diversity

Results from our previous experiments show that while some feature types yield similar accuracies independently, such as those in Table 5, 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 was recently investigated by Malmasi and Cahill (2015) who used Yule's Q coefficient of association as one approach to measuring the degree of diversity between features. They applied this method to English data from the TOEFL11 corpus and in this section we expand this evaluation to include other languages. We begin by comparing the approach to an ablative analysis and describing the Q coefficient.

[36] For example, an NLI system with 70 per cent accuracy against an Oracle baseline of 80 per cent is relatively better compared to one with 74 per cent accuracy against an Oracle baseline of 93 per cent.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

An ablation study is a common approach in machine learning that aims to measure the contribution of each feature in a multicomponent 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.[37] 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.* (2006) 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.

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) predictions made by that learner. Each classifier C_i produces a result vector $y_i = [y_{i,1}, \ldots, y_{i,N}]$ for a set of N documents where $y_{i,j} = 1$ if C_i correctly classifies the jth 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_iCorrect	N^{11}	N^{10}
C_iWrong	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 follows:

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

This distribution-free association measure[38] is based on taking the products of the diagonal cell frequencies and calculating the ratio of their difference and sum.[39] Q

[37] Other variations exist, e.g. Richardson *et al.* (2006) and Wellner *et al.* (2006)
[38] We also note that this is equivalent to the 2×2 version of Goodman and Kruskal's gamma measure for ordinal variables.
[39] Division by zero is possible here, see Bakeman and Quera (2011, 115) for more details.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

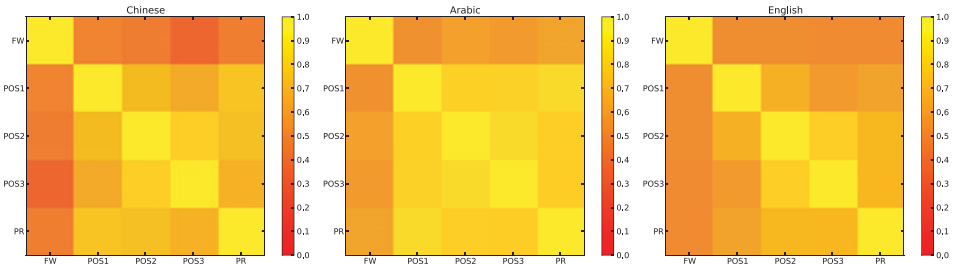

Fig. 12. (Colour online) The Q coefficient matrices of five features for Chinese (l), Arabic (m) and English (r). The matrices are displayed as heat maps. POS 1/2/3: POS uni/bi/trigrams, FW: Function words, PR: Production rules.

ranges between −1 and +1, where −1 signifies negative association, 0 indicates no association (independence) and +1 means perfect positive correlation (dependence).

In this experiment, our classifiers 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.

11.1 Results

We calculate the Q coefficient for our largest dataset, Chinese, using all five features listed in Table 5. 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 12. We did not find a negative correlation between any of our features.

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 12. This seems to suggest that these correlation patterns may hold cross-lingually.

To test the validity of these results, we re-run the Chinese experiment from Section 6, this time combining the top three features with the lowest average Q coefficient, weighted by their classification error.[40] These features are Function words, Production rules and POS trigrams and combining them yields an accuracy of 70.7 per cent, compared to 70.6 per cent 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.

[40] For feature i, this is calculated as $\bar{Q}_i \times (1 - \text{Accuracy}_i)$; lower values suggest higher accuracy and diversity.

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

11.2 Discussion

Such analyses can help us better understand the linguistic properties of the features and guide interpretation of the results. This analysis can be used to examine the orthogonality or complementarity between features, particularly those of the same type. One such example would be a comparison between two syntactic features, context-free grammar production rules and grammatical dependencies, both of which are based on parsing and known to be useful for NLI. The measure is most useful when comparing features with similar individual performance to identify those with the highest diversity. As shown by our results, this can be utilized for finding combinations of diverse and high performing features.

This information 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 research can be applied to that end. By selecting fewer but less redundant features, it should be possible to build simpler models with equal, if not better, performance.

Another promising application is in a cross-corpus setting where a model is trained on one corpus and tested on a different one. This approach has been applied in NLI using several L2 English corpora (Tetreault *et al.* 2012; Ionescu, Popescu and Cahill 2014; Malmasi and Dras 2015b). In one such study, Brooke and Hirst (2012b) investigated the utility of a standard feature set across several corpora and compared their cross-corpus performance, similar to our within-corpus study in Experiment I. In this context, the feature diversity measures can be applied in a cross-corpus setting to see if the same patterns found here are also present. This can help us better understand interfeature correlations when they are applied across differing genres, domains and registers and how this may differ from the same correlations within a single corpus.

An important direction for future research is the expansion of the current analysis to more languages and features, as the required resources become available. This could help identify specific feature correlations that are present across languages.

The analysis could also be expanded within a single corpus. The largest number of NLI features have been explored for English and they include, *inter alia*, Tree Substitution Grammar fragments, Brown clusters, Adaptor Grammars, Dependencies, spelling errors and *n*-gram language models. A logical extension of this line of inquiry is the application of this method to study the levels of interdependence and redundancy between these myriad features. Such an analysis has not been attempted to date and could yield insightful findings for NLI and SLA research.

12 General discussion and conclusion

The present study has examined a number of different issues from a cross-lingual perspective, making a number of novel contributions to NLI research. Using up to six language 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

https://doi.org/10.1017/S1351324915000406 Published online by Cambridge University Press

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 at approximately similar rates 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 per cent accuracy. We also looked at some issues that have not previously been addressed for NLI. Experiment V also presented a new type of NLI error analysis aimed at calculating the upper limits of classification accuracy in this task, something not done to date. Our feature diversity analysis in Section 11 presented an approach for estimating the diversity and dependence of linguistic features.

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.

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 L2, 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 L2 transfer effects during L3 acquisition have been a recent focus in CLI 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 L2 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 CLI 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 (Dewaele 1998; Williams and Hammarberg 1998; De Angelis 2005).

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 CLI 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. Such ensemble classifiers have also proven to be useful for related tasks such as dialect identification (Malmasi and Dras 2015a; Malmasi *et al.* 2015a). The methods described in our feature diversity analysis from Section 11 can help guide the selection of diverse features to reduce redundancy in the classifier committee.

Acknowledgements

We would like to thank the anonymous reviewers for their extensive and constructive comments. We would also like to thank Ilmari Ivaska and Kirsti Siitonen making the Finnish learner data available. We also thank Anne Ife for providing the Spanish learner corpus.

References

Abbasi, A., and Chen, H. 2005. Applying authorship analysis to extremist-group Web forum messages. *IEEE Intelligent Systems* **20**(5): 67–75.

Abuhakema, G., Faraj, R., Feldman, A., and Fitzpatrick, E. 2008. Annotating an Arabic learner corpus for error. In *LREC*, Marrakech, Morocco.

Alfaifi, A., and Atwell, E. 2013. Arabic learner corpus v1: a new resource for arabic language research. In *Proceedings of the Second Workshop on Arabic Corpus Linguistics, 2013*, Leeds, UK.

Alfaifi, A., Atwell, E., and Hedaya, I. 2014. 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.

Bakeman, R., and Quera, V. 2011. *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge, England, UK: Cambridge University Press.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. 2013. TOEFL11: a corpus of non-native english. Technical Report, Educational Testing Service.

Branch, M. 2009. Finnish. In B. Comrie (eds.), *The World's Major Languages*, pp. 497–518. London: Routledge.

Brooke, J., and Hirst, G. 2011. Native language detection with 'cheap' learner corpora. In *Conference of Learner Corpus Research (LCR2011)*, Presses universitaires de Louvain, Louvain-la-Neuve, Belgium.

Brooke, J., and Hirst, G. 2012a. Measuring interlanguage: native language identification with L1-influence metrics. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, pp. 779–784.

Brooke, J., and Hirst, G. 2012b. Robust, lexicalized native language identification. In *Proceedings of COLING 2012*, The COLING 2012 Organizing Committee, Mumbai, India, pp. 391–408.

Carreras, X., Chao, I., Padró, L., and Padró, M. 2004. FreeLing: an open-source suite of language analyzers. In *LREC*, Lisbon, Portugal.

Cenoz, J., Hufeisen, B. and Jessner, U. 2001. *Cross-Linguistic Influence in Third Language Acquisition: Psycholinguistic Perspectives*, vol. 31, Multilingual Matters.

Charniak, E., and Johnson, M. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the Meeting Assoc. Computat. Linguistics (ACL)*, Ann Arbor, Michigan, pp. 173–180.

Chen, M. 2013. 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): 418–442.

Chen, J., Wang, C., and Cai, J. 2010. *Teaching and Learning Chinese: Issues and Perspectives*. IAP.

Comrie, B. 1989. *Language Universals and Linguistic Typology*, 3rd ed. Chicago, IL, US: University of Chicago Press.

Corino, E. 2008. VALICO: an online corpus of learning varieties of the Italian language. In *Proceedings of the 2nd Colloquium on Lesser Used Languages and Computer Linguistics*, Bolzano, Italy, pp. 117–133.

Coulthard, M., and Johnson, A. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. London: Routledge.

De Angelis, G. 2005. Multilingualism and non-native lexical transfer: an identification problem. *International Journal of Multilingualism* **2**(1): 1–25.

Dewaele, J.-M. 1998. Lexical inventions: French interlanguage as L2 versus L3. *Applied Linguistics* **19**(4): 471–490.

Diab, M. 2009. Second generation AMIRA tools for Arabic processing: fast and robust tokenization, POS tagging, and base phrase chunking. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Diéz-Bedmar, M. B., and Papp, S. 2008. The use of the English article system by Chinese and Spanish learners. *Language and Computers* **66**(1): 147–176.

Ellis, R. 2008. *The Study of Second Language Acquisition*, 2nd ed. Oxford University Press, Oxford, UK.

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

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. 2008. LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research* **9**(Aug): 1871–1874.

Gamon, M., Chodorow, M., Leacock, C., and Tetreault, J. 2013. Using learner corpora for automatic error detection and correction. In N. Ballier, A. Díaz-Negrillo, and P. Thompson (eds.), *Automatic Treatment and Analysis of Learner Corpus Data*, pages 127–150. Studies in Corpus Linguistics. Amsterdam, The Netherlands: John Benjamins Publishing Company.

Garside, R. 1987. The CLAWS word-tagging system. In *The computational analysis of English: A Corpus Based Approach*. London: Longman.

Gass, S. M., and Selinker, L. 2008. *Second Language Acquisition: An Introductory Course*. New York: Routledge.

Gibbons, J. 2003. Forensic linguistics: an introduction to language in the justice system. https://books.google.com.au/books/about/Forensic_Linguistics.html?id=hVPsw4DWjGAC.

Gibbons, J., and Prakasam, V. 2004. *Language in the Law*. Telangana: Orient Blackswan.

Granger, S. 1994. The learner corpus: a revolution in applied linguistics. *English Today* **10**(03): 25–33.

Granger, S. 2003. 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.

Granger, S. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching. *Corpora and Language Teaching* **33**: 13–32.

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

Grant, T. 2007. Quantifying evidence in forensic authorship analysis. *International Journal of Speech Language and the Law* **14**(1): 1–25.
Green, J. N. 2009. Spanish. In B. Comrie (eds.), *The world's Major Languages*, pp. 197–216. London: Routledge.
Greene, B. B., and Rubin, G. M. 1971. Automated grammatical tagging of English. Department of Linguistics, Brown University, 1971, pages 306.
Guo, Y., and Beckett, G. H. 2007. The hegemony of english as a global language: reclaiming local knowledge and culture in china. *Convergence* **40**(1–2): 117–132.
Gyawali, B., Ramirez, G., and Solorio, T. 2013. Native language identification: a simple n-gram based approach. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Atlanta, Georgia, pp. 224–231.
Habash, N. Y. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies* **3**(1): 1–187.
Hawkins, J. A. 2009. German. In B. Comrie (eds.), *The World's Major Languages*, pp. 86–109. London: Routledge.
Horst, M., White, J., and Bell, P. 2010. First and second language knowledge in the language classroom. *International Journal of bilingualism* **14**: 331–349.
Iggesen, O. A. 2013. *Number of Cases*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Ionescu, R. T., Popescu, M., and Cahill, A. 2014. Can characters reveal your native language? A language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar.
Ivaska, I. 2014. The corpus of advanced learner Finnish (LAS2): database and toolkit to study academic learner Finnish. *Apples - Journal of Applied Language Studies* **8**(3): 21–38.
Jarvis, S., Bestgen, Y., and Pepper, S. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Atlanta, Georgia, pp. 111–118.
Jarvis, S., and Crossley, S. (eds.) 2012. *Approaching Language Transfer Through Text Classification: Explorations in the Detection-based Approach*. Bristol, UK: Multilingual Matters.
Kaye, A. S. 2009. Arabic. In B. Comrie (eds.), *The World's Major Languages*, pp. 560–577. London: Routledge.
Kochmar, E. 2011. *Identification of a Writer's Native Language by Error Analysis*. MPhil thesis, Cambridge, UK: University of Cambridge.
Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, Montreal, Quebec, Canada, vol. 14, pp. 1137–1145.
Koppel, M., and Schler, J. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the 21st International Conference on Machine Learning*, ICML '04, ACM, New York, NY, USA, p. 62. ISBN 1-58113-838-5.
Koppel, M., Schler, J., and Zigdon, K. 2005a. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*, vol. 3495 pp. 209–217. Lecture Notes in Computer Science. Berlin: Springer-Verlag.
Koppel, M., Schler, J., and Zigdon, K. 2005b. Determining an author's native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ACM, Chicago, IL, pp. 624–628.
Kuncheva, L. I. 2002. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(2): 281–286.
Kuncheva, L. I., Bezdek, J. C., and Duin, R. P. 2001. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition* **34**(2): 299–314.

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., and Duin, R. P. 2003. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications* **6**(1): 22–31.

Lahiri, S., and Mihalcea, R. 2013. Using n-gram and word network features for native language identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Atlanta, Georgia, pp. 251–259.

Lam, L. 2000. Classifier combinations: implementations and theoretical issues. In *Multiple classifier systems*, pages 77–86. Berlin: Springer.

Laufer, B., and Girsai, N. 2008. Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics* **29**(4): 694–716.

Levy, R., and Manning, C. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Sapporo, Japan, pp. 439–446.

Li, C. N., and Thompson, S. A. 2009. Chinese. In B. Comrie (eds.), *The world's Major Languages*, pp. 703–723. London: Routledge.

Lightbown, P. M. 2000. Anniversary article. Classroom SLA research and second language teaching. *Applied Linguistics* **21**(4): 431–462.

Lozano, C. 2009. CEDEL2: Corpus escrito del Español L2. In C. M. e. a. Bretones Callejas (eds.), *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*, pp. 197–212. Almería: Universidad de Almería.

Lozanó, C., and Mendikoetxea, A. 2010. Interface conditions on postverbal subjects: a corpus study of L2 English. *Bilingualism: Language and Cognition* **13**(4): 475–497.

Lozano, C., and Mendikoetxea, A. 2013. Learner corpora and second language acquisition: the design and collection of CEDEL2. In N. Ballier, A. Díaz-Negrillo, & P. Thompson (eds.) *Automatic Treatment and Analysis of Learner Corpus Data*. pp. 65–100. Amsterdam: John Benjamin.

Malmasi, S., and Cahill, A. 2015. Measuring feature diversity in native language identification. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 49–55. Association for Computational Linguistics, Denver, Colorado.

Malmasi, S., and Dras, M. 2014a. Chinese native language identification. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden.

Malmasi, S., and Dras, M. 2014b. Language transfer hypotheses with linear SVM weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar.

Malmasi, S., and Dras, M. 2015a. Automatic language identification for Persian and Dari texts. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, Bali, Indonesia, pp. 59–64.

Malmasi, S., and Dras, M. 2015b. 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)*, Association for Computational Linguistics, Denver, CO, USA, pp. 1403–1409.

Malmasi, S., Refaee, E., and Dras, M. 2015a. Arabic dialect identification using a parallel multidialectal corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, Bali, Indonesia, pp. 209–217.

Malmasi, S., Tetreault, J., and Dras, M. 2015b. Oracle and human baselines for native language identification. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Denver, Colorado, pp. 172–178.

Malmasi, S., Wong, S.-M. J., and Dras, M. 2013. NLI shared task 2013: MQ submission. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Atlanta, Georgia, pp. 124–133.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. 2014. The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, USA, pp. 55–60.

Mansouri, F. 2005. Agreement morphology in Arabic as a second language: typological features and their processing implications. In *Cross-linguistic aspects of Processability Theory*, John Benjamins Publishing, Amsterdam, pp. 117–153.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. 1993. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics* **19**(2): 313–330.

McDonald, R. T. *et al.* 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 92–97.

McMenamin, G. R. 2002. *Forensic linguistics: Advances in Forensic Stylistics*. FL, USA: CRC Press.

Ministry of Labour 2006. Hallituksen maahanmuuttopoliittinen ohjelma. *Tyhallinnon julkaisu 371*.

Monroe, W., Green, S., and Manning, C. D. 2014. Word segmentation of informal Arabic with domain adaptation. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, USA.

Mosteller, F., and Wallace, D. L. 1964. *Inference and Disputed Authorship: The Federalist*. Reading, MA, US: Addison-Wesley.

Murphy, S. 2005. Second language transfer during third language acquisition. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics* **3**(2003): 1–21.

Nesselhauf, N. 2004. Learner corpora and their potential for language teaching. In Sinclair, John (ed.), *How to Use Corpora in Language Teaching*, Amsterdam: Benjamins, pp. 125–152.

Nielsen, H. L. 1997. On acquisition order of agreement procedures in Arabic learner language. *Al-Arabiyya* **30**: 49–93. Georgetown University Press. http://www.jstor.org/stable/43192775.

Nieminen, T. 2009. Becoming a new Finn through language: non-native English-speaking immigrants' views on integrating into Finnish society. https://jyx.jyu.fi/dspace/handle/123456789/22355.

Ortega, L. 2009. *Understanding Second Language Acquisition*. Oxford, UK: Hodder Education.

Padró, L., and Stanilovsky, E. 2012. 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.

Perkins, R. 2014. *Linguistic Identifiers of L1 Persian Speakers Writing in English: NLID for Authorship Analysis*. Ph.D. thesis, Birmingham, UK : Aston University.

Petrov, S., Das, D. and McDonald, R. 2012. A universal part-of-speech tagset. In N. C. C. Chair, K. Choukri, T. Declerck, M. U. Doan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey. ISBN 978-2-9517408-7-7.

Pirkola, A. 2001. Morphological typology of languages for IR. *Journal of Documentation* **57**(3): 330–348.

Richards, J. C., and Rodgers, T. S. 2014. *Approaches and Methods in Language Teaching*. Cambridge, UK: Cambridge University Press.

Richardson, M., Prakash, A., and Brill, E. 2006. Beyond PageRank: machine learning for static ranking. In *Proceedings of the 15th International Conference on World Wide Web*, ACM, New York, pp. 707–715.

Ringbom, H. 2001. Lexical transfer in L3 production. In Cenoz and Jessner (2001), pp. 59–68.

Rose, H., and Carson, L. 2014. Introduction. *Language Learning in Higher Education* **4**(2): 257–269.

Rozovskaya, A., and Roth, D. 2011. 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*, Portland, Oregon, USA, pp. 924–933.

Ryding, K. C. 2013. Teaching Arabic in the United States. In K. M. Wahba, Z. A. Taha, and L. England (eds.), *Handbook for Arabic Language Teaching Professionals in the 21st Century*. London: Routledge, pp. 13–20.

Sampson, G. 1993. The SUSANNE corpus. *ICAME Journal* **17**(125127): 116.

Schachter, J. 1974. An error in error analysis. *Language Learning* **24**(2): 205–214.

Schiller, A., Teufel, S., and Thielen, C. 1995. Guidelines für das Tagging deutscher Textcorpora mit STTS. *Manuscript, Universities of Stuttgart and Tübingen*.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. 2001. Estimating the support of a high-dimensional distribution. *Neural Computation* **13**(7): 1443–1471.

Siemen, P., Lüdeling, A., and Müller, F. H. 2006. FALKO-ein fehlerannotiertes Lernerkorpus des Deutschen. *Proceedings of Konvens 2006*. Konstanz, Germany.

Siitonen, K. 2014. 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.

Swanson, B., and Charniak, E. 2012. Native language detection with tree substitution grammars. In *Proceedings of the Meeting Assoc. Computat. Linguistics (ACL)*, Jeju Island, South Korea, pp. 193–197.

Swanson, B., and Charniak, E. 2013. 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*, Association for Computational Linguistics, Atlanta, Georgia, pp. 85–94.

Swanson, B., and Charniak, E. 2014. Data driven language transfer hypotheses. In *EACL 2014*, Gothenburg, Sweden, pp. 169.

Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics* **1**(2013): 1–12.

Tetreault, J., Blanchard, D., and Cahill, A. 2013. A report on the first native language identification shared task. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Atlanta, Georgia, pp. 48–57.

Tetreault, J., Blanchard, D., Cahill, A., and Chodorow, M. 2012. Native tongues, lost and found: resources and empirical evaluations in native language identification. In *Proceedings of COLING 2012*, The COLING 2012 Organizing Committee, Mumbai, India, pp. 2585–2602.

Tinsley, T. 2013. *Languages: the state of the nation: demand and supply of language skills in the UK*. London, UK: British Academy.

Tomokiyo, L. M., and Jones, R. 2001. You're not from round here, are you? Naive Bayes detection of non-native utterance text. In *Proceedings of the 2nd North American Chapter of the Association for Computational Linguistics*, NAACL '01, Pittsburgh, PA, USA, pp. 239–246.

Torney, R., Vamplew, P., and Yearwood, J. 2012. Using psycholinguistic features for profiling first language of authors. *Journal of the American Society for Information Science and Technology* **63**(6): 1256–1269.

Tsung, L., and Cruickshank, K. 2011. *Teaching and Learning Chinese in Global Contexts: CFL Worldwide*. London, UK: Bloomsbury Publishing.

Tsur, O., and Rappoport, A. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on*

Cognitive Aspects of Computational Language Acquisition, Association for Computational Linguistics, Prague, Czech Republic, pp. 9–16.

Vergyri, D., Kirchhoff, K., Duh, K., and Stolcke, A. 2004. Morphology-based language modeling for Arabic speech recognition. In *INTERSPEECH*, Jeju Island, Korea, vol. 4, pp. 2245–2248.

Vincent, N. 2009. Italian. In B. Comrie (eds.), *The World's Major Languages*, pp. 233–252. London: Routledge.

Wahba, K. M., Taha, Z. A., and England, L. 2013. *Handbook for Arabic language Teaching Professionals in the 21st Century*. London: Routledge.

Wang, M., Malmasi, S., and Huang, M. 2015. The Jinan chinese learner corpus. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Denver, Colorado, pp. 118–123.

Warrens, M. J. 2008. On association coefficients for 2× 2 tables and properties that do not depend on the marginal distributions. *Psychometrika* **73**(4): 777–789.

Wellner, B., Pustejovsky, J., Havasi, C., Rumshisky, A., and Sauri, R. 2009. Classification of discourse coherence relations: an exploratory study using multiple knowledge sources. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, Association for Computational Linguistics, Sydney, Australia, pp. 117–125.

Williams, S., and Hammarberg, B. 1998. Language switches in L3 production: implications for a polyglot speaking model. *Applied Linguistics* **19**(3): 295–333.

Wong, K.-F., Li, W., Xu, R., and Zhang, Z.-S. 2009. Introduction to Chinese natural language processing. *Synthesis Lectures on Human Language Technologies* **2**(1): 1–148.

Wong, S.-M. J., and Dras, M. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA)*, Sydney, Australia, pp. 53–61.

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

Wong, S.-M. J., Dras, M., and Johnson, M. 2012. 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*, EMNLP-CoNLL '12, Jeju Island, Korea, pp. 699–709.

Wozniak, M., and Zmyslony, M. 2010. Designing fusers on the basis of discriminants–evolutionary and neural methods of training. In *Hybrid Artificial Intelligence Systems*, pp. 590–597. Berlin: Springer.

Wu, C.-Y., Lai, P.-H., Liu, Y., and Ng, V. 2013. Simple yet powerful native language identification on TOEFL11. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Atlanta, Georgia, pp. 152–156.

Xia, F. 2000. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0).

Yule, G. U. 1912. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society* **75**(6): 579–652.

Zhao, H., and Huang, J. 2010. 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.