Native Dialect Influence Detection (NDID): Differentiating between Mexican and Peninsular L1 Spanish in L2 English

Andrea Mojedano Batel, Mitchell Abrams & Piotr Pęzik

Aston Institute for Forensic Linguistics, Aston University, UK

https://doi.org/10.21747/21833745/lanlaw/9_1a6

Abstract. *The current investigation addresses a vital lacuna in forensic authorship studies, and more concretely, in Native Language Influence Detection (NLID) research: narrowing down a speaker's native dialect instead of only their native language (L1), which might not be enough when carrying out sociolinguistic profiling tasks. Native Dialect Influence Detection (NDID), the focus of our study, can thus greatly aid at the investigative level. We approach this topic by providing a comprehensive analysis of linguistic features that serve to identify two non-contact dialects of L1 Spanish (i.e., Mexican and Peninsular varieties) when dealing with data written in L2 English, which come from Tripadvisor. Our main aim is to investigate if an author's L2 features can point to their L1 native dialect, rather than only to their native language. Findings point to L1 dialectal transfer of punctuation signs, adjectives of affect, and intensifiers: these linguistic features, even when expressed in an L2, show a culturally bound use. Additionally, we implemented an automatic classifier that achieved an accuracy of 69% in categorizing test data, using only linguistic features that have explanatory power and can aid linguistic theory. This is key for explainability in the forensic context, which Native Language Identification (NLI) studies tend to neglect (Kingston 2019). Results show that L1 Spanish dialects can be differentiated by analyzing L2 English text, pointing to NDID as a fertile approach for narrowing down candidate L1 dialects of a language when analyzing L2 data.*

Keywords: *Native dialect influence detection, Native language influence detection, Authorship analysis, Language variety identification, Spanish.*

Resumo. *A investigação atual aborda uma lacuna essencial nos estudos de autoria forense, mais concretamente na investigação sobre Deteção de Influência da Língua Materna (NLID): afunilar a análise do dialeto materno de um falante em vez de focar apenas a sua língua materna (L1), o que pode não ser suficiente quando se realiza tarefas de caracterização sociolinguística. A Deteção da Influência da Língua Materna (NDID), o enfoque do nosso estudo, poderá assim auxiliar*

significativamente a investigação. Abordamos este tema realizando uma análise abrangente das características linguísticas usadas para identificar dois dialetos sem contacto do espanhol L1 (isto é, variedades mexicanas e peninsulares) em dados escritos em inglês L2 provenientes do Tripadvisor. O nosso principal objetivo é investigar se as características L2 de um autor podem apontar para o seu dialeto nativo L1, e não apenas para a sua língua materna. Os resultados apontam para a transferência dialetal L1 de pontuação, adjetivos afetivos e intensificadores. Estas características linguísticas, mesmo quando expressas numa L2, revelam uma utilização culturalmente marcada. Adicionalmente, implementámos um classificador automático que alcançou uma precisão de 69% na categorização dos dados do teste utilizando apenas características linguísticas que possuem capacidade explicativa e podem contribuir para fundamentar a teoria linguística, fundamental para fornecer fundamentações em contexto forense, o que os estudos de Identificação Língua Materna (NLI) tendem a negligenciar (Kingston 2019). Os resultados mostram que os dialetos espanhóis L1 podem ser diferenciados através da análise de texto L2 em inglês, apontando a NDID como uma abordagem fértil para reduzir os dialetos candidatos L1 de uma língua ao analisar os dados L2.

Palavras-chave: *Deteção de influência de dialeto nativo, Deteção de influência de língua materna, Análise de autoria, Identificação de variedade linguística, Espanhol.*

Native Dialect Influence Detection (NDID): Differentiating between Mexican and Peninsular L1 Spanish in L2 English

This study provides a comprehensive description of linguistic features that serve to identify non-contact dialects of a native language (L1) when dealing with second language (L2) data, carrying out bottom-up quantitative and qualitative analyses of token n-grams and part-of-speech (POS) n-grams, leaving aside other feature analyses for future studies. For this purpose, we have collected a corpus of texts written in L2 English by L1 Mexican Spanish (MxSp) and L1 Peninsular Spanish (PenSp) authors on the Tripadvisor website.[1] The study's main aim is to investigate if an author's L2 features can point to their L1 native dialect, rather than only to their native language. We term this specific subdomain of Native Language Influence Detection (NLID) *Native Dialect Influence Detection* (*NDID*). The current study draws on research from the fields of Second Language Acquisition (SLA), forensic linguistics, contact linguistics, sociolinguistics, and computational linguistics.

Addressing NDID lacunae will aid the forensic linguist in providing "evidence for criminal and civil investigations and courtroom disputes," which is one of the main aims of the field (Coulthard *et al.* 2010: 529). The overarching goal of NDID studies is to estimate group belonging rather than to identify individuals. In the specific case of Spanish, with 483 million native speakers worldwide (Fernández Vítores 2020), knowing that a speaker's L1 is Spanish when dealing with L2 data might not be enough; NDID helps narrow down the possible native and/or dominant regional dialect of a speaker. This type of sociolinguistic profiling, while unlikely to be admissible as evidence in the UK judicial and court systems (Grant 2008; Perkins and Grant 2013), can greatly aid at the investigative level, benefitting law enforcement agencies or organizations, as well as intelligence agencies dealing with non-native speakers (Perkins 2015). Casework in NLID is seldom published due to the very nature of intelligence work (Perkins 2015), but it is

possible to conduct NLID research taking into consideration intelligence applications by collaborating with relevant and interested agencies and departments (e.g., Grant *et al.* (2010)), and the same could be said of NDID as a sub-domain of NLID. Spanish, the L1 explored in this study, is the third most used language on the Web as of March 2020, with almost 364 million Spanish speakers using the Internet, which corresponds to 7.9% of all Internet users (http://www.internetworldstats.com). Hence, there is an abundance of Spanish data on the web; it must come as no surprise that many of these speakers might also express themselves in other languages (e.g., English), and these data should be explored from a forensic linguistic viewpoint.

The present research lies within the framework of forensic authorship analysis, and more concretely, it takes a Native Language Influence Detection (NLID) approach. NLID seeks to reveal an author's native language (L1) from common interlanguage phenomena that may occur in any non-native language they write in, which can be their second language, but also their third, fourth, etc. We will refer to this non-native language output in the present study as L2 in order to simplify matters; however, forensic linguists must consider that a speaker can have more than one native language, as the majority of the planet's population is bilingual (Thomason 2001).

Within the field of computational linguistics, the term *Native Language Identification* (*NLI*) is used, referring to an often wholly automatic computational method for the identification of the language or languages that likely influenced a text or group of texts, based in machine learning and data mining, and introduced by Koppel *et al.* (2005). Since then, a variety of classifier approaches and feature sets have been proposed, the most popular including word n-grams, part-of-speech (POS) n-grams, character n-grams, function words, dependency features, and spelling features. These were all features exploited in the 2013 NLI Shared Task (Tetreault *et al.* 2013). Additionally, production rules were also a successful feature in Wong and Dras (2011).

A frequent problem with NLI studies is that they tend to neglect the features used to classify texts (Kingston 2019). While features for these computational tasks are leveraged mainly to achieve high classification accuracy, they are key for explainability in the forensic context. Some studies, however, do provide a light analysis of linguistically motivated features in NLI tasks (e.g., Koppel *et al.* (2005); Jiang *et al.* (2014); Goldin *et al.* (2018)).

In the following section, we will address relevant studies in the fields of computational linguistics and forensic linguistics in order to provide the reader with information on what has been done already and what still needs to be tackled in the field of NDID.

Previous Accounts of Dialect Identification, with an Emphasis on Spanish

Research dealing with dialect identification lies within the interface between computational linguistics (more specifically, within Natural Language Processing, or NLP) and authorship profiling. The term *Language Variety Identification* (*LVI*) refers to "labelling the text in a native language (e.g., Spanish, Portuguese, English) with its specific variation (e.g., Argentina, Chile, Mexico, Peru, Spain; Brazil, Portugal; UK, US)" (Rangel *et al.* 2016: 1). LVI investigates a dialect of a language written in an L1 and therefore does not directly pertain to NLID studies, yet shares with NDID research the commonality of focusing on the differentiation of dialects of a particular language.

LVI studies point to certain automated features being useful to discern among dialects. Sadat *et al.* (2014) used character n-gram features to discriminate between six different Arabic dialects, obtaining accuracies between 70%-80%. Zampieri and Gebrekidan-Gebre (2012) looked at automatic classification of written Brazilian and European Portuguese, using a character n-gram model and a word n-gram model; their results showed an accuracy of over 90%. Maier and Gómez-Rodríguez (2014) investigated LVI in Spanish tweets from Argentina, Chile, Colombia, Mexico, and Spain. They obtained 60%-70% accuracy by applying language modeling techniques that combined four types of features: character n-gram frequency profiles, character n-gram language models, Lempel-Ziv-Welch compression, and syllable-based language profiles. Rangel *et al.* (2016) used Low Dimensionality Representation (LDR) to differentiate between Argentinian, Chilean, Mexican, Peruvian, and Peninsular Spanish. Instead of selecting the most frequent n-grams, Rangel *et al.* (2016) assigned higher weights to the most discriminative words in each dialect. Their results showed that Peninsular Spanish was the easiest dialect to discriminate and that Latin American varieties were closer to each other than to Peninsular Spanish, and therefore, it was more difficult to differentiate among them. Nevertheless, the highest precisions were attained for Mexico and Peru. While LVI studies help us gauge what kinds of computational features can aid in the task of differentiating among dialects, a frequent problem with these studies, much like with NLI studies, is that they generally neglect the linguistic features used to classify texts and that results from character n-gram analysis cannot usually be explained from a linguistic viewpoint.

The only study, to the best of our knowledge, that has tackled identification of a speaker's native dialect rather than simply their native language by looking at L2 data is Kingston (2019), who analyzed L2 English data from L1 speakers of three French dialects: Metropolitan (i.e., from France), Canadian, and Maghrebi. Because the primary goal of Kingston's study was to "designate potentially distinguishing features of native speakers of different dialects of the same language" (p.26), comparing three dialects where two of them were in contact situations (Canadian French with English and Maghrebi French with Arabic and Tamazight) made it potentially easier to identify dialects vis-a-vis identifying non-contact dialects of a language, which is the aim of our study.

Kingston (2019) secondary aim was to detect the extent to which machine learning classification mirrored the human analyst's findings. She used a decision tree classifier (J48) in the Waikato Environment for Knowledge Analysis (WEKA). Kingston concluded that purely computational methods could achieve very good classification accuracy, but that there was a need for a more inclusive approach taking into account linguistic typology, sociolinguistic data, and semantic and syntactic analysis. The present investigation seeks to address some of these lacunae.

In the present study we chose Peninsular and Mexican Spanish dialects for data analysis because previous investigations have shown that Latin American dialects are closer to each other than to Peninsular (Lipski 2012; Rangel *et al.* 2016). If our results show that we cannot tell the difference between Spaniards' and Mexicans' L2 English output, it might prove more difficult to differentiate English L2 outputs among Latin American dialects. Additionally, Mexican Spanish is the Spanish dialect with the largest number of native speakers in the world (121 million speakers, according to Fernández Vítores (2020)). Being able to examine if there are any differences between Peninsular and Mex-

ican Spanish dialects with regard to speakers' L2 English output will help pave the way for further studies considering other Spanish dialects.

Preference for British or American English as L2 Input and Output

While, to the best of our knowledge, there is no official information as to which dialects of English Mexican and Peninsular Spanish speakers are exposed to, findings from (Caraker 2016) show that several teachers from Central Spain believed Peninsular Spanish students to be more exposed to British English (BrE). Meanwhile, Despagne (2010) reported that in Mexico, television programs are more than 50% in English and the dialect of English that these media mainly use is the American dialect. This points to Mexican Spanish speakers possibly having more influence from American English (AmE) than British English in their L2 English input and output.

Sue Garton (personal communication, July, 2020) believes that most European countries favor standard British English (if there is such a thing, as she notes), while Latin America would be more oriented toward a US variety. However, Garton does not think there are many official policies, and much is likely to depend on the coursebook that is used and the dialect that the teacher speaks. The difference between American and British English when it comes to L2 English instruction is vital, as it could possibly bring different grammatical, lexical, and orthographical choices that we need to consider when carrying out our analysis.

Data and Methodology

This section offers a detailed account of the corpora, the extraction of adequate tokens, and the methodology used in applying both a computational linguistics approach and a forensic linguistics approach to account for dialectal differences in the L2 English output of L1 Peninsular and Mexican Spanish speakers.

The data we present in this study are corpus-based. They come from Tripadvisor, an open internet source, and the genre pertains to computer-mediated communication (CMC). Features that characterize grammar in electronic communication vary systematically across languages, contexts, users, and technological modes (Herring 2012). Herring (2012) notes that non-standard orthography is common in CMC, where users can be lax about orthographic norms. Bieswanger (2008) demonstrated through a systematic comparison of English and German text messages (SMS) that Britons and Germans favor different shortening strategies when texting and that these strategies are used in different proportions. In terms of our Tripadvisor data, by the very nature of the website, the entries are evaluative (which restaurants, hotels, and locations authors liked or disliked), and the entries' main function is communicative, with suggestions for, and requests to, other users. Our findings can thus be transferable to forensic linguistic contexts, because forensic texts can be evaluative and often display requests and/or demands.

We compiled four corpora, listed below:

- A corpus with training L2 English data produced by L1 Mexican Spanish speakers (word count: 37,500, number of entries: 514, number of authors: 38, average entry length: 73 words)
- A corpus with training L2 English data produced by L1 Peninsular Spanish speakers (word count: 40,602, number of entries: 504, number of authors: 37, average entry length: 81 words)

- A corpus with test L2 English data produced by L1 Mexican Spanish speakers (word count: 6,481, number of entries: 53, number of authors: 6, average entry length: 122 words)
- A corpus with test L2 English data produced by L1 Peninsular Spanish speakers (word count: 4,797, number of entries: 50, number of authors: 9, average entry length: 96 words)

We are aware that corpora sizes are small, but it was very difficult to make them larger due to their specialized nature: further studies could build upon this initial research with data from other sources. To ensure that the authors were native speakers of their respective dialects, we first needed to find posts asserting their linguistic identity, such as "I am / I'm Mexican / Spanish / from Mexico / from Spain," "Being from Mexico / Spain," "I live in Mexico / Spain," etcetera. Moreover, Tripadvisor users had to have a geolocator in their account so that we could be more certain they lived in either Mexico or Spain; authors who said that they were from either of these two countries but were geolocated in a different country were discarded because of language contact and language identity issues. Additionally, because we were examining L2 English data, we had to make sure that the participants spoke Spanish as an L1. To do this, we checked their posts to make sure they had written at least one of them in Spanish and that in said post(s), they showed native-like command of Spanish.[2]

For each of the two training corpora, we took the most recent 25 entries per author when there were more than 25 entries for that specific person; if there were less than 25 entries, we took them all; the minimum number of entries per author in the training corpora was one. For each of the two test corpora, we took the five most recent entries per author when there were more than five entries for that specific person; if there were less than five entries, we took them all; the minimum number of entries per author in the training corpora was one.

While we chose texts from all over Mexico, we only chose texts from areas in Spain where Spanish is not in contact with other languages (i.e., Central and Southern Spain) with the purpose of narrowing dialectal variation. It can be argued that Mexican Spanish is in contact with indigenous languages all over the Mexican territory: while this is certainly true, only 6% of the population (approximately six million people) speak an indigenous language, according to the Commission for the Development of Indigenous Peoples (CDI) and the National Institute of Indigenous Languages (INALI). Thus, it is improbable that entries from Mexican Spanish speaking authors come from a Mexican contact dialect.

Before entries in the target and reference corpora were processed for tokenization, tagging, and analysis, the data went through a pre-processing step for cleaning and collecting user post frequency. We kept the raw text for analysis since extra-linguistic conventions of the written online medium (namely, parentheses, quotation marks, ellipses, punctuation, emoticons, and website links) can potentially be influenced by a person's native dialect. In this same vein, we refrained from normalizing the language through lemmatizing, correcting misspellings, and removing capitalization. One exception we made was to replace various instances of accented *í* with *i* for English words, as these were attributed to keyboard mistakes rather than an individual's language, and including these characters in the data would have significantly altered token frequencies. In the final pre-processing step, we collected user post counts to check for feature distri-

bution, so that a feature under analysis was not over-attributed to any one individual in the target corpus.

As a starting point in investigating native dialect influence, we created a list of token n-grams and part-of-speech (POS) n-grams ranked by their keyness. The term *token* in our study refers to words and non-words in our corpora (e.g., word forms, punctuation, digits, abbreviations, etc.) and serves as the basis of our analysis. Even parts-of-speech were leveraged to reveal token realizations and shallow syntactic structures. An English language tokenizer was applied internally through SketchEngine (Kilgarriff *et al.* 2014) before n-gram analysis. An *n-gram* is a contiguous sequence of N words or tokens. For the present study in particular, we collected one to six grams. With these units for analysis, we ranked them by keyness as a way of operationalizing consistency and distinctiveness in the target corpora.

We approached keyness from a corpus linguistic perspective: tokens, POS sequences, and grammatical patterns that are key occur in texts with outstanding frequency as compared with a reference corpus. We believe that keywords are not key by themselves, but because they are frequently used in particular lexical combinations or grammatical patterns, which makes them attractive units of analysis to explore NDID tasks. This study takes two approaches to identify keywords in texts, namely, a keyness metric through SketchEngine software (Kilgarriff *et al.* 2014), and Burrows' Zeta (Burrows 2007) to assess the dispersion of style markers used by authors represented in the corpus. Results were first collected separately and then compared later for analysis, confirming that one approach had not overlooked any important patterns. The keyness metric in SketchEngine software uses the simple maths method (Kilgarriff, 2009) to show how many times a given word is more frequent in a target corpus (corpus 1) than in a reference corpus (corpus 2). This is a simple metric, yet effective enough for our relatively small dataset, comprised of posts of a relatively similar size and of a consistent genre. Alongside keyness scores, we extracted the frequencies to see how much more frequent an n-gram was in one corpus over another. Burrow's Zeta, on the other hand, was calculated as the difference between the proportion of n-word segments in which a given marker such as a word n-gram or POS-gram was found in the Mexican and Peninsular Spanish corpora. This approach was meant to operationalize the more abstract notion of style marker consistency in our study.

The main motivation for extracting token and POS n-grams as linguistic features for analysis, and excluding others (e.g., character n-grams), is that they have explanatory power which can aid both linguistic theory and forensic research. Token n-grams in particular are ideal for this study and reveal noteworthy patterns. Additionally, in computational research, token n-grams have been proven to achieve high accuracy in related NLID and LVI tasks. In such analyses, data scarcity (notably with our smaller target corpus size) becomes a problem, because linguistic items appear in low frequencies, especially when focusing on higher token n-gram sequences (such as bigrams and trigrams). However, even with this limitation in mind, it is possible to identify similarities and differences between observed linguistic patterns in our corpora. We also handled this sparsity issue by leveraging the hidden categories of part-of-speech (POS) tags to reveal more distinct patterns. On the level of morphosyntactic annotation, we analyzed recurrent POS n-gram patterns, which allowed us to capture grammatical patterns that were not directly observable. For our data in particular, we used two methods

for tagging the corpora; one that used the built-in SketchEngine software part-of-speech tagger with the English Penn Treebank tagset, and another that utilizes the POS tagger available in the Spacy (Honnibal and Montani 2015).

Results and Analysis

In the present section, we examine our results, taking a mixed-methods approach to their analysis. First, we discuss token n-grams, divided into intensifiers, adjectives of affect, quantifiers, contracted forms, and punctuation marks. We then discuss POS n-grams, and more concretely, bigrams and trigrams. Finally, we offer a description and analysis of our classification task, providing researchers with a starting point for NDID classification. When reporting descriptive statistics, we include both raw and relative frequencies (unless otherwise stated) for token n-grams and raw frequencies for POS n-grams. Relative frequencies are provided through percentages, as our corpora are small in size.

Token N-grams

Intensifiers

Intensifiers can be classified in intensives and downtoners, and both for AmE and BrE, the most frequent collocation pattern for intensifiers (72%) is with adjectival heads (Bäcklund 1973).[3] In our study, raw and relative frequencies of intensifiers shed light on cross-dialectal semantic and pragma-linguistic differences. In this subsection we discuss the difference in frequency and use of intensives *quite* and *really* and then examine the downtoner *a (little) bit*.

The intensive *quite*. In line with what was expected of Peninsular Spanish speakers, the adverb *quite*, more frequently used in BrE than in AmE (395.14 times per million, according to BNC, vs. 182.71 times per million, according to COCA, respectively), appears in the Peninsular Spanish corpus almost thrice as often than in the Mexican Spanish corpus (N = 58, 0.14% vs. N = 21, 0.05%, respectively).[4] A log-likelihood (LL) test of significance indicated that the difference in use of the intensive *quite* between corpora was significant ($p < 0.05$). Both sets of authors used this adverb most frequently in partially lexically-filled constructions consisting of [be + quite + AdjP/DP], as in (1); yet, the difference in frequency of use can probably be attributed to the assumption that Spaniards are more exposed to a British variety of English, whereas Mexicans generally learn an American English variety.

(1a) *since the broschure [sic]* ***was quite complicated*** *to understand* (PenSp corpus)
(1b) *Marbella* ***is quite a big place*** (PenSp corpus)
(1c) *getting there* ***is quite easy*** (MxSp corpus)

The intensive *really*. In line with what was expected of Mexican Spanish authors, the adverb *really*, more frequently used in AmE than in BrE (902.34 times per million vs. 458.10 times per million, respectively), appears in the Mexican Spanish corpus almost twice as often as in the Peninsular Spanish corpus (151 times vs. 89 times, respectively), mirroring AmE rates of use. A log-likelihood (LL) test of significance indicated that the difference in use of the intensive *really* between corpora was significant ($p < 0.05$). Moreover, while *really* modifies adjectives in a similar way in both MxSp and PenSp, Mexicans also use *really* to modify adverbs, something seldom done by Spaniards in our corpus (N = 9, 0.02% vs. N = 4, >0.01%, respectively), as shown in (2).

(2) *Everything was so amazing and delicious, and served* ***really fast*** (MxSp corpus)

The downtoner *a (little) bit.* Spanish makes ample use of the diminutive morpheme *-ito/-ita*, whose role extends beyond the expression of small size, giving emotional tone to words (Travis 2004). Thus, Travis (2004) argues, not only the lexicon, but also the morphosyntax of a language reflects the cultural values of its speakers. Furthermore, she notes that extensive use of the diminutive is evidence of a cultural value associated with good feelings, a claim that Wierzbicka (1984, 1992) has also made with regard to Russian and Polish. Concerning Spanish dialects, Company Company (2002) has shown that there is wider use of the diminutive in Mexican Spanish compared to Peninsular Spanish, both in terms of overall frequency and range of pragmatic functions. A similar result is provided by Reynoso Noverón (2001), who found that in a corpus of written and oral narratives, Mexicans used the diminutive to encode small size just 28% of the time, whereas Spaniards used it for this same reason 58% of the time: these results show that Peninsular participants in Reynoso Noverón's study tended to use the diminutive morpheme chiefly for encoding small size, whereas Mexican participants generally used it in a wider range of pragmatic functions, and not predominantly for encoding small size.

Standard Spanish makes use of the quantifier *poco* 'a bit' to convey the idea of a small amount. *Poco* can appear with the diminutive morpheme *ito/ita*, i.e., *poquito/poquita*; these lexemes are attested in both Mexican and Peninsular Spanish. Moreover, Mexican Spanish speakers use the quantifiers *tantito/tantita* 'a bit / a little bit' to express the idea of a small amount.

English doesn't encode the diminutive as a morpheme, and Spanish speakers have to make use of other means to express small size in their L2 English output. When examining the construction *a (little) bit* in our corpora —where parentheses indicate optionality— a first observation is that both dialects show, for the most part, similar frequencies (N = 45, 0.11% for Peninsular Spanish and N = 31, 0.08% for Mexican Spanish) and distributional patterns (a log-likelihood (LL) test of significance indicated that the difference in use of the construction *a (little) bit* was not significant). There is one important difference, however: [*be + a (little) bit*] constructions, namely, *a (little) bit* constructions which co-appear with the copula *to be*. This is the most frequent construction in which *a (little) bit* appears in both corpora, but its behavior is quite different depending on the L1 Spanish dialect of the speaker.

The construction [*be + a (little) bit*] shows a raw frequency of six instances in the Mexican corpus and of 24 instances in the Peninsular Spanish corpus. In both, speakers generally use the adverbial form *a (little) bit* to try to downplay a negative trait of something that they're describing, as in (3).

(3a) *The service* ***was a bit slow*** (MxSp corpus)
(3b) *The décor* ***could be a little bit more modern*** (MxSp corpus)
(3c) *january [sic]* ***can be a bit rainy*** (PenSp corpus)
(3d) *The shower* ***is a little bit old*** *but it goes together with the style of the parador [sic].* (PenSp corpus)

When analyzing the differences in use of the [*be + a (little) bit*] construction in both corpora, we found that, similarly to what takes place in Mexican and Peninsular Spanish in general, there is wider use of a diminutive form (in this case, *little*) in the Mexican

Spanish L2 English output compared to the Peninsular Spanish L2 English output (N = 4, 66.6% of all [*be + a (little) bit*] constructions in MxSp vs. N = 3, 12.5% of all [*be + a (little) bit*] constructions in PenSp). We tested for the difference in the frequency of the [*be + a little bit*] construction between corpora with a log-likelihood (LL) test of significance: the LL is 4.65, showing the difference to be significant at the $p < 0.05$ level (or at the 95% level). That is, although Spaniards produce [*be + a (little) bit*] more than Mexicans, they overwhelmingly do so in its [*be + a bit*] form, while Mexicans mainly produce the [*be + a little bit*] construction to convey the same idea of downplaying a negative trait.

Moreover, when analyzing the form *a little bit* by itself (that is, without the copula *to be*), the use of this construction followed patterns found in L1 Spanish: there was a wider use of the construction with the diminutive form *little* in the Mexican Spanish corpus (N = 12) compared to the Peninsular Spanish corpus (N = 8) in terms of range of pragmatic functions: Mexicans, apart from using it to downplay negative emotions —the only function that appeared in Peninsular Spanish use—, also used it once to downplay requests (4a), once to encode positive feelings (4b), and twice to encode quantity, as in (4c).

(4a) *Hi guys* ***I'm looking for a little bit of information*** *regarding a trip me and a few friends are planning* (MxSp corpus)
(4b) *A* ***little bit of everything****: Tulum & Xel-ha* (MxSp corpus)
(4c) *You do need to know* ***a little bit of history*** *of France* (MxSp corpus)

Adjectives of Affect: *Good, Great, Excellent,* and *Amazing*

When examining the lemma *good* (that is, in its different forms *good*, *better*, *best*) we see that it is more frequent in the Mexican Spanish corpus than in the Peninsular Spanish one (N = 392, 1.05% vs. N = 355, 0.87%). We tested for the difference in the frequency of the lemma *good* between corpora with a log-likelihood (LL) test of significance: the LL is 5.95, showing the difference to be significant at the $p < 0.05$ level (or at the 95% level). In example (5) below, we can gauge Mexican authors' use of the adjective *best* in context.

(5a) *This was by far our* ***best meal*** *in Hanoi* (MxSp corpus)
(5b) *This is truly one of the* ***best beaches*** *in Mexico* (MxSp corpus)
(5c) ***Best dinner*** *in town!* (MxSp corpus)

The adjective *best*, like other adjectives we discuss in this section, is a stance marker of affect. Precht (2000, 2003a,b) showed that stance expression is systematically different across cultures and contended that we tend to identify and stereotype people based on their stance use. As Precht (2003a) notes, there has been little research carried out regarding dialectal differences in stance, at least in English. To the best of our knowledge, this seems to be the case for Spanish varieties as well. Her (2003a) study comparing BrE and AmE showed that both varieties had a very similar conversational stance across semantic categories; however, British speakers displayed lower frequencies than Americans for affect markers that expressed emotion and for emphatics, while having higher frequencies for modal verbs. From these results, it appears that Americans directly express emotion more frequently than Britons.

The same could be said for Mexicans vis-a-vis Spaniards in terms of their stance expression: Mexican speakers seem to directly express their emotion, at least in their L2 English output, more often than Peninsular Spanish speakers. Besides producing the lemma *good* in a variety of forms more often than their Spaniard counterparts, Mexicans

also show a significantly higher use ($p < 0.05$) of the following adjectives of affect: *great* (N = 207, 0.55% vs. N = 132, 0.33%), *excellent* (N = 73, 0.19% vs. N = 37, 0.09%), and *amazing* (N = 51, 0.13% vs. N = 36, 0.09%). Figure 1 provides a visual breakdown of the difference in frequencies of adjectives of affect between corpora.

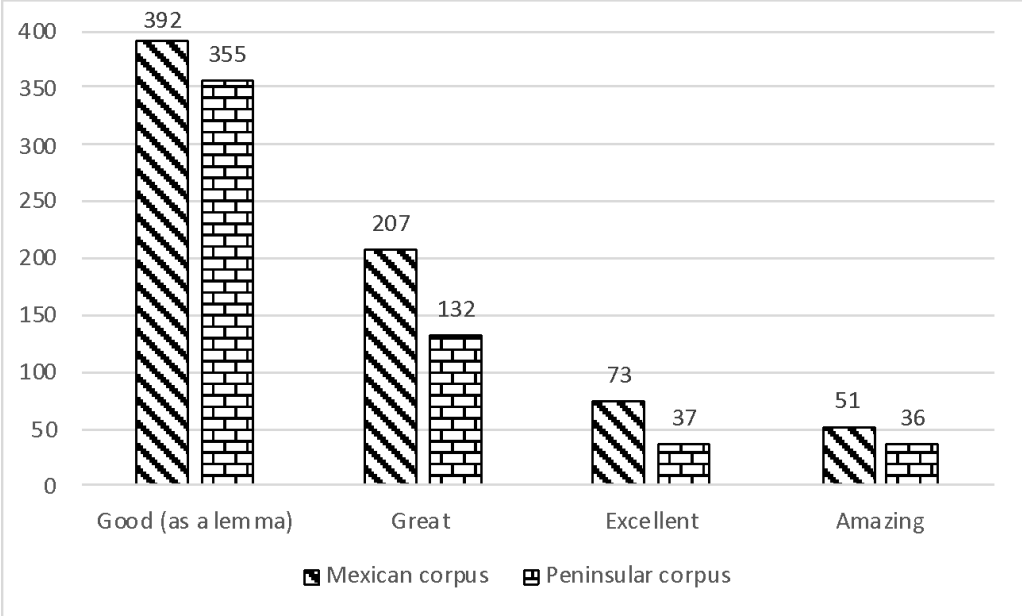

Figure 1. Frequency of Use of Adjectives of Affect by Corpus)

Example (6) below shows different uses of adjectives of affect in both corpora.

(6a) *Belsaita is right and her* ***suggestions are great*** (PenSp corpus)
(6b) *Enjoy! Coba and Tulum* ***are real [sic] great*** (MxSp corpus)
(6c) *A stylish place with* ***excellent Peruvian cuisine****.* (MxSp corpus)
(6d) *The* ***starters are excellent*** *and the pasta with sea food is really impressive.* (PenSp corpus)
(6e) *all of a sudden you are surrounded by an [sic]* ***amazing architecture*** *with a lot of history* (PenSp corpus)
(6f) ***Everything was so amazing*** *and delicious, and served really fast* (MxSp corpus)

One construction with adjectives of affect, however, is only used by Spaniards, with no instances in the Mexican corpus: the token bigram *very good* (7).

(7) ***Very good*** *Chinese food in El Puig at reasonable price.* (PenSp corpus)

The significant difference in frequency of use of adjectives of affect in Spaniards' and Mexicans' L2 English output ($p < 0.05$ in all cases) could be explained by Mexican speakers being more influenced by the way Americans express emotion through affect markers. Yet, as far as we know, there is a lacuna of research concerning Spanish dialectal differences in stance markers, especially with regard to stance markers of affect. Thus, it could also very well be that Mexicans express emotion more frequently than Spaniards: more research is needed in order to draw conclusions.

There is another possible reason for the differences we found in frequency of use of adjectives of affect in both corpora, which does not necessarily invalidate our previous

explanation: Spaniards in the corpus may have a more limited repertoire of use of L2 English adjectives of affect, and consequently make more use of Spanish calques when writing in English. The construction *very good* is a fine example, as it is a word by word translation of Sp. *muy bueno*. Furthermore, *very good* could be understood as a compositional phrase. Compositional phrases, following Snider and Arnon (2012), originate in the grammar while non-compositional phrases originate in the lexicon and are stored together with their idiosyncratic syntactic and semantic features. Since they are derived in a predictable way, compositional phrases do not need to be stored in the lexicon (Snider and Arnon 2012), and thus, should require less processing effort from the speaker who produces them. This could point to Peninsular Spanish speakers in the corpus having a lesser command of English than their Mexican counterparts.

Quantifiers[1]

The augmentative *many* in the construction *many thanks*. The construction *many thanks* is found exclusively in the Peninsular Spanish corpus. Mexican authors either thanked someone on Tripadvisor by writing *thanks* (N = 61, 0.16%) —optionally accompanied by *a lot* or *a ton*—or, to a lesser extent, by writing *thank you* (N = 35, 0.09%) —sometimes co-occurring with *so much* after it. Meanwhile, Spaniards primarily used constructions with *thanks* (N = 77, 0.19%), while *thank you* only occurred on 15 occasions (0.03%) in the Peninsular Spanish corpus. We tested for the difference in frequency of the word *thanks* and the construction *thank you* between corpora with a log-likelihood (LL) test of significance, which indicated that the difference in use of the construction *thank you* to be significant ($p < 0.05$) while the difference in use of the word *thanks* was not significant. Figure 2 provides a visual breakdown of the difference in frequencies of gratitude statements between corpora.

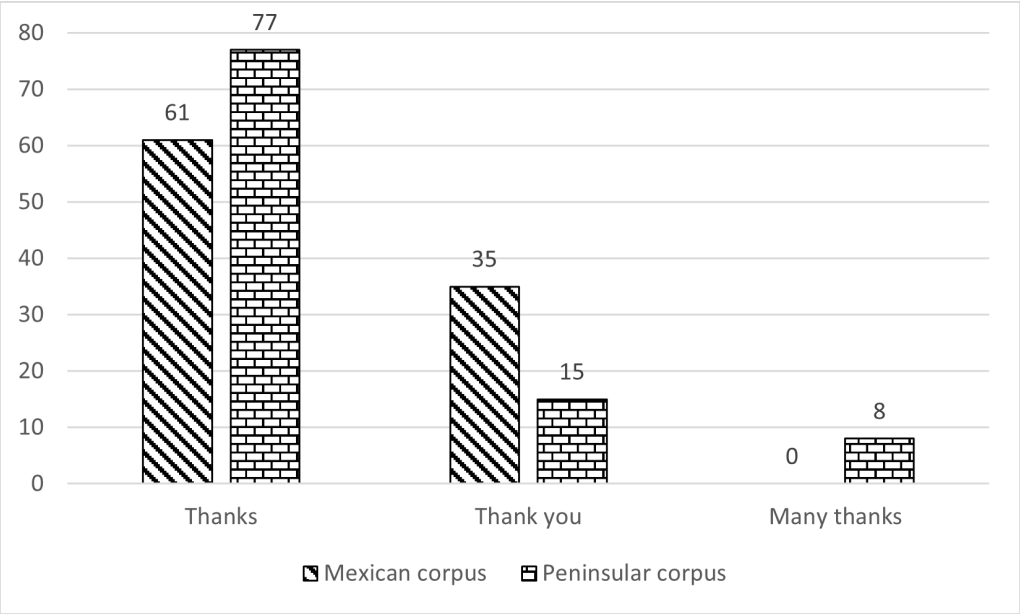

Figure 2. Frequency of Use of Gratitude Statements by Corpora)

[1]The diminutive has already been discussed as a downtoner in the previous subsection.

Of the *thanks* constructions found in the Peninsular Spanish data, six (0.01%) co-appeared with *a lot*, as in *thanks a lot*, and more notably, 16 (0.04%) co-appeared with *many*, as in *many thanks* (8).

> (8) *Could you give any idea??.* ***many thanks*** *again. greetings from Madrid.* (PenSp corpus)

Many thanks, which can be understood as a compositional phrase, only occurs in the Peninsular corpus: the difference in quantities between corpora proved significant ($p < 0.05$). As previously mentioned, compositional phrases do not need to be stored in the lexicon (Snider and Arnon 2012), and thus, should require less processing effort from the speaker who produces them. This, like other examples in the present study (e.g., the construction *very good*) point to Peninsular Spanish speakers in the corpus possessing a lesser command of English than their Mexican counterparts. It could also be argued that Spaniards in the corpus have understood the word *thanks* as Spanish *gracias* (English *thank you*), and because usually this *gracias* appears with the intensifier *muchas* (English *many*) before it, they have made a calque.

This theory is reinforced by a quick comparison of English terms to thank someone. The use of *many thanks* by Spaniards cannot be explained by British English influence alone, as the BNC (BNC Consortium 2007) only shows 2.36 instances per million words (Table 1). Furthermore, in both native dialects of English, *thank you* is the preferred way of thanking someone, whereas in both our L2 English corpora, authors preferred the use of *thanks*.

Term	**BNC**	**COCA**
	Frequency per million words	**Frequency per million words**
Many thanks	2.36	1
Thanks a lot	1.47	4.93
Thanks	62.97	206.27
Thank you	95.55	284.35

Table 1. English Terms Used to Thank Someone: Frequencies

Crystal (2008: 111) defines a contraction as "the process or result of phonologically reducing a linguistic form so that it comes to be attached to an adjacent linguistic form," e.g., *I've* from *I have*, and *haven't* from *have not*, as well as the *wanna* contraction. Both verb contractions and *not*-contractions, following Young (2015), are well represented in standard written texts across various registers and are familiar to non-native speakers (NNSs) of English. Biber (1987), using nine genres from the Lancaster-Oslo-Bergen (LOB) and Brown standardized corpora to compare British and American writing, found systematic differences in the frequencies of contractions: all American genres displayed a markedly higher frequency for this feature than the same British genres, although Biber coded for auxiliary and negative contractions together and unfortunately did not provide quantitative evidence comparing the two contraction strategies. Other more specific analyses of contraction, such as (Algeo 2006), have shown, however, that contraction with the verb *have* is more frequent in BrE. Yaeger-Dror (1997) provided evidence that negative contraction is conditioned by interactional and other register variables, and through her analysis of pragmatic and morphological interpretation of negatives, she

found that negative contraction and auxiliary contraction must be distinguished from each other.

The acquisition of English auxiliary contraction and deletion by second-language learners has been a largely unexplored question (Samar 2003). Research on related topics has shown that copula/auxiliary *be* is one of the first morphemes acquired by both child and adult L2 learners (Krashen 1977; Lightbown 1987), yet, very little is known about how these learners acquire variable contraction of auxiliaries. Odlin (1978) examined the acquisition of English contractions by six Spanish native speakers from Mexico who were studying English as a Second Language (ESL) in Texas. Odlin's results suggested evolutionary stages in the acquisition of contractions characteristic of a certain level of proficiency, where contraction frequency generally correlated with general proficiency (i.e., more proficient students used contracted forms more frequently and with a greater variety of words than less proficient students). The stages of English contraction acquisition would begin with word classes with a single member (such as *it* or *that*), following with other pronouns and *that* and *there*, finally evolving to more complex noun phrases and locatives. The problem with Odlin's stages is that the differences between "word classes with a single member" and "other pronouns" are not clear, and thus, we could not apply his continuum analysis to our results. A further problem is that the acquisition of negative contracted forms (separate from auxiliary contraction) in English as a foreign language is an understudied topic. In the present study we strive to provide a description of linguistic features that serve to identify non-contact dialects of a native language (L1) when dealing with second language (L2) data; taking into consideration human, topic and time constraints on analysis capacity, we compared contraction usage rates between dialects without comparing the two contraction strategies. Future studies in NDID could provide a separate qualitative analysis of negative contraction and auxiliary contraction.

We now turn to our own data, taking previous investigations into account in order to draw certain conclusions. Table 2 allows us to compare contraction usage rates by dialect.

Feature	**Mexican L1 Spanish Corpus**		**Peninsular L1 Spanish Corpus**	
	Number of authors using feature	Segment proportion	Number of authors using feature	Segment proportion
" 's "	23	0.086	non-significant	non-significant
" n't "	24	0.071	21	0.076
" 'm "	15	0.032	20	0.037
" 're "	10	0.013	non-significant	non-significant
" 've "	9	0.012	14	0.017
" 'd "	9	0.01	non-significant	non-significant

Table 2. Usage rates of contracted forms by dialect

Table 2 shows six significant contracted forms in the Mexican Spanish corpus, with three of these contracted forms also being significant in the Peninsular Spanish corpus. We see similar rates of usage and of segment proportions for the significant contracted forms in both corpora. Additionally, while the word *is* was the sixth most frequent unigram found in the Mexican Spanish corpus, this same word was the most frequent unigram in

the Peninsular Spanish corpus, possibly pointing to the comparably lower rates in which contractions occur with the copula *be* in Peninsular authors' L2 English output.

Bigram results in both corpora are also enlightening. Data shows, in general, a more frequent use of contracted forms over their respective full form alternatives in the Mexican Spanish corpus (as examples, *don't* occurs 64 times and *do not*, six; *it's* occurs 62 times and *it is*, 50). In contrast, in the Peninsular Spanish corpus, *it is* occurs 131 times, *It is*, 52 times, and *it's*, 32. Trigrams in our data are no different in this matter. Within the twenty most frequent trigrams in the Peninsular Spanish corpus, we found ten full forms, such as *is not the*, *I do not*, or *I am from*, and zero contracted forms; meanwhile, in the Mexican Spanish corpus, within the twenty most frequent trigrams, we encountered three contracted forms, such as *I don't* and *I'm from*, and only full forms (*if you are* and *would like to*).

These findings demonstrate that Mexican authors make a wider and more varied use of contractions than Peninsular authors. Results also possibly point to Spaniards in the corpus having lower proficiency English levels than Mexicans; having said that, we cannot discount the fact that because Spaniards have probably more input from British English—a dialect with considerably less use of contractions in written registers (Biber 1987) —, their reduced use of contractions could be partly due to British English influence.

Punctuation

Within the punctuation inventory, there are three functional classes (Bredel (2008, 2011), cited in Busch (2021)): syntactic signs, communicative signs, and scanning signs. To guide the grammatical parsing process, syntactic signs are used (i.e., period, comma, colon, and semicolon); to mediate the relationship between writer and reader, communicative signs are used (i.e., exclamation mark, question mark, quotation mark, and parentheses); to indicate that information to decode the message is missing and must be interpreted by the reader, scanning signs are used (i.e., hyphen, apostrophe, and ellipsis dots). As Busch (2021) notes, in some forms of CMC, the use of punctuation signs by collaborative writers in their interactions (such as the ones we find in our TripAdvisor data) acts as contextualization cues, in order to indicate the interpretation of utterances and guide sequential progress.

Previous studies have found that the use of emoji, another form of non-word tokens in CMC, is strongly impacted by cultural background (Gibson *et al.* 2018). Additionally, politeness is culturally bound (Terkourafi 2011). Accordingly, it would be expected for punctuation signs —especially for communicative and scanning signs—to show cross-cultural and cross-dialectal variation.

There are important differences in frequency of use of certain communicative and scanning signs between the Peninsular and Mexican Spanish data in our corpora, shown in Table 3.

Specifically, Mexicans in our data use ellipsis dots (both in its standard three-dot form [...] as well as in repeated forms with four or more dots [....]) more than Spaniards. In (9a), ellipsis dots are used in order to change topic; in (9b), as an ending to the text; in (9c-d), as a way to add suspense and introduce an unexpected outcome, be it positive (9c) or negative (9d); and to delve deeper into a topic (9e). Sometimes, as in (9f), an author can use ellipsis dots in more than one occasion, for different pragma-linguistic

Feature	Mexican Spanish Corpus	Peninsular Spanish Corpus
…	4,04	3,5
…. (4 or more dots)	548	21
!	5,295	4,259
!!!	320	274
?	4,131	2,213
???	182	84
mean	2,419	1,725
SD	2312.60	1871.46

Table 3. Relative frequencies of punctuation signs per million tokens

functions (i.e., as a way to add suspense and introduce an unexpected outcome, and to change topics).

(9a) *Hi, I'm from Guadalajara, and a huge Chivas fan. The games usually start at 07:00 pm. I hope this gives some answers...*
TICKETS: You can buy them during the week (Monday-Friday) at different locations across the city. The day of the game tickets are sold at the stadium until they finish [...] (MxSp corpus)
(9b) *This show is a must when you're in New York. It has it all: great plot, outstanding music, girfted [sic] singers...everything! I definitely recommend Les Mis...* (MxSp corpus)
(9c) *Looking for a great sushi place... This is it!*
We stumbled upon this by chance, but had a mayor surprise (MxSp corpus)
(9d) *Happy birthday... except for the check* (MxSp corpus)
(9e) *Yeah, it's worse due to the dry season, heat and sun...somehow it makes ozone stick around longer* (MxSp corpus)
(9f) *We went there a few days ago because we were visiting the Oceanografic. We just went for a walk to see the buidings and the pools ... Amazing! It looks like you were in a futuristic city ... There are also activities if you go with your kids ...*
You cannot miss this if you visit Valencia! (PenSp corpus)

Moreover, Mexicans in our corpus make use of exclamation marks and question marks, both single and iterated, more often than Spaniards. Example (10) shows some punctuation mark usage within our corpora.

(10a) *The Castle in Chichen.... you can climb it !* (MxSp corpus)
(10b) *The food is very good and the service is very attentive.*
We couldn't try one of their specialties, apparently, they cook Camembert in a very special and delicious way ... next time, hopefully!!
The only thing is that the service is not the fastest in the world... so, take it easy and enjoy the views!!
Congratulations Ca'María team!! (PenSp corpus)
(10c) *I studied at Bristol university for a semester 26 years ago....and I'm coming back!!!* (PenSp corpus)

This analysis of results requires further qualitative examination to clarify the possible reasoning for punctuation mark usage in our data, taking into account the context in which these punctuation marks were used and analyzing them as "contextualization cues" (Busch 2021: 3) that allow language users to "signal contextual presuppositions"

(Gumperz 1982: 131). The literature tackling politeness and requests in the Spanish speaking world points to possible misunderstandings between speakers of different dialects, where (most) issues seem to be found in differing politeness systems, especially between Latin Americans and Spaniards, as Spaniards are perceived as very direct and abrupt by Latin Americans, and Latin Americans are understood by Spaniards to be too formal sometimes (see Márquez-Reiter (2002), for further discussion on the topic).[5] Márquez-Reiter (2002) provides a possible explanation as to why Spaniards are regarded as more direct than Latin Americans, namely, Spaniards' lack of tentativeness when carrying out a request compared to Uruguayan speakers.

The use of punctuation signs in our data signals more tentativeness and more explicit politeness devices by Mexican users, in line with previous research. This seems to be achieved by Mexicans' higher use of ellipsis dots and exclamation and question marks. Our findings align with Bieswanger (2008), who indicated that in CMC, users favor different strategies in different languages and that these strategies are used in different proportions. Our findings also align with Gibson *et al.* (2018), showing that non-word tokens in CMC are impacted by cultural background.

Part-of-Speech (POS) N-grams

We must consider syntactic complexity as an important measure of L2 proficiency (Larsen-Freeman 1978; Lu 2011) that can assist us in gauging differences between authors, and possibly, between different L1 dialects when producing L2 output. Ortega (2015) defines syntactic complexity (also called *syntactic maturity* or *linguistic complexity*) as "the range and the sophistication of grammatical resources exhibited in language production" (p. 82). Biber and colleagues (Biber *et al.* 2011, 2014), when examining formal academic registers, indicated that spoken registers showed grammatical complexity in their clausal elaboration (i.e., subordination with high-frequency mental verbs such as *think*, *know*, and *say*) while written registers displayed grammatical complexity through phrasal compression, and specifically through complex noun phrases and complex phrases. Tyler and Evans (2001) showed that prepositions are significant markers of linguistic complexity, as more advanced L2 learners start being able to use lower-frequency prepositions and prepositional phrases and use these in more idiomatic extended senses. Kyle and Crossley (2018: 334) suggest using fine-grained indices of syntactic complexity, such as phrasal complexity indices, as they "may provide a clearer understanding of the relationship between syntactic complexity and L2 writing" than coarse-grained indices that measure complexity at the clause or sentence level.

Taking these studies into consideration, we now proceed to analyze syntactic complexity through POS bigrams and trigrams. POS n-grams that were most distinctive, that is, that were ranked as distinctive by both the keyness metric and by the Burrows' Zeta metric (versus only appearing as distinctive in one of the two metric lists) are the ones that we discuss below.

POS Bigrams

POS bigrams provide information about differences in how Spaniards and Mexicans express themselves. The construction [VERB ADP], realized as phrasal verbs, is distinctive of the Mexican Spanish corpus. The presence of phrasal verbs points to Mexican authors in our corpora possessing higher lexical sophistication in English than Spaniard authors. The most frequent examples in our data include the verbs *to go to, to look for, to be from,*

to stay in, to go to, to get to, to feel like, to be in, to ask for, and *to live in*. Phrasal verbs are said to be inherently difficult for learners of English as a second language (ESL) to master, with research indicating that ESL students often avoid them and frequently make mistakes in their attempt to use them (Bronshteyn and Gustafson 2015).

Another construction distinctive of the Mexican Spanish corpus is the bigram [NOUN NOUN], where most frequent examples point to Mexican authors' ability to use complex nouns such as *day trip, front desk, observation deck, bus station, science museum, train tickets, train schedule*, and *wine list.* As Biber and colleagues (Biber *et al.* 2011, 2014) have argued, the main source of linguistic complexification in written academic language is phrasal compression through complex noun phrases and complex phrases. It seems that we can transpose this feature of linguistic complexity not only to academic writing, but also to more informal registers, such as the CMC genre of Tripadvisor entries.

POS Trigrams

A distinctive POS trigram in our Peninsular Spanish data is the construction [ADJ NOUN PUNCT], with one of its realizations being *Many thanks!* (N = 5). As has been discussed previously, this phrase is possibly a calque from Sp. *Muchas gracias*, and as a compositional phrase, it does not need to be stored in the lexicon (Snider and Arnon 2012), requiring less processing effort from the speaker who produces it. Possibly, Peninsular Spanish authors in our corpus have a less idiomatic command of English than Mexican Spanish authors, relying more on simpler structures and trying to compositionally define many constructions. This can also be perceived by Spaniard authors' distinctive use of the [AUX ADV ADJ] construction, with which, instead of describing a noun with an adverb and an adjective before it (e.g., *a very*[ADV] *good*[ADJ] *restaurant*[NOUN]), they described a noun in a compositional manner (the restaurant[NOUN] is[AUX] very[ADV] good [ADJ]), as in (11).

(11a) *The staff* ***is very good****, helpful and friendly.* (PenSp)
(11b) *Price* ***was very good*** *too* (PenSp)
(11c) *The food is great, the service* ***is very attentive*** *and the atmosphere is calm and very pleasant* (PenSp)

The [PRON AUX VERB] construction was the fourth most frequent trigram in both the Peninsular Spanish (N = 415) and Mexican Spanish (N = 360) corpora. The most frequent examples of this construction in the Mexican Spanish corpus, however, show a wider variety of verbs, verb tenses and pronominal persons than its Peninsular Spanish counterpart, as can be seen in Table 4 below. More specifically, Mexican Spanish —in opposition to Peninsular Spanish—includes the *be going to* future tense (*you are going to*), use of *should*, use of contracted forms (*'m*), and use of the first-person plural pronoun *we*.

Classification task

To supplement our corpus analysis, we investigated the discriminatory power of grammatical features in an automatic classification task. Specifically, we tested whether a particular combination of features could predict if a Tripadvisor post was produced by an L1 speaker of Peninsular Spanish or Mexican Spanish. Our primary goal was to understand how linguistic features can aid the forensic linguist in discerning against dialects, not to tweak the classification model to achieve the highest accuracy possible. Because linguistic features are key for explainability in the forensic context, we avoided using

Peninsular Spanish Corpus	**Raw Frequency**	**Mexican Spanish Corpus**	**Raw Frequency**
You will find	12	You should try	9
You can find	12	I would like	8
You are looking	10	You would expect	7
I would recommend	10	I'm planning	7
I would like	9	You can go	6
I would say	8	You can take	6
You can see	7	I was thinking	5
You can take	6	I would recommend	5
You can go	6	We would like	5
You can have	5	You are going	5

Table 4. The [PRON AUX VERB] construction by dialect

opaque features such as character n-grams or word embeddings (Mikolov *et al.* 2013) for our classification task.

We trained and developed on the target corpora and saved the classifier weights for our held-out test set. We implemented a logistic regression model in Python with the Scikit-learn library (Pedregosa *et al.* 2011), using one-hot encoding as a vectorization method for each feature. One-hot encoding accounts for the presence of a feature in a text, notwithstanding its frequency. This approach was a good fit for our experiment, as texts were not very long; longer texts might require frequency encoding.

Since we had a binary classification task, we considered a 50% accuracy result to be our chance baseline. We also reported on precision, recall[6], and F-1 (the harmonic mean of the precision and recall score) to offer a holistic evaluation. Since the F-1 score is the weighted average of the precision and recall score, it punishes extreme scores and offers a better understanding of the performance of the models.[7]

Our first baseline models leveraged token unigrams and token bigrams as features. Since these are typical features in text classification, they offer a starting point for seeing how well we can classify the texts without sophisticated feature engineering. This model considered all tokens in the text as features—either a single token context or a two-token context. For instance, for the text "*you should try*", unigrams represent the individual tokens (*you, should, try*) and bigrams the two-token sequences (*you should, should try*). Consequently, this allowed the classifier to consider tokens beyond the scope of our analysis (e.g., content words) and encoded other linguistic patterns that we did not examine in detail (e.g., short token sequences or shorter syntactic patterns).

As displayed in Table 5, both unigram and bigram classifier variants yielded an accuracy of 65%, an improvement on the chance baseline (50%), demonstrating that there are linguistic features in the text that can discriminate between the two L1 dialects in L2 English. The unigram and bigram classifiers show identical precision, recall, F-1, and accuracy scores because both models are similarly influenced by the presence of the same tokens—either the presence of a single token (unigram) or a two token sequence (bigram)—, but how these tokens are represented does not seem to make a difference for predicting a label in our data. Despite the predictive power of these baseline models,

content words can be misleading features. For example, many words in our corpus give away the speaker's L1 dialect by mentioning places such as *Mexico*, *Madrid*, *Mexico City*, *Spain*, *Monterrey*, *Sevilla*, *Cancun*, and *Guadalajara*. The classifier weighs these words heavily, yet such features are topic dependent.

Next, we turned to feature engineering with the linguistic categories from our qualitative analysis, namely, punctuation, adjectives of affect, intensifiers, and contracted forms. We also considered the performance of combining these features to improve classification accuracy. We did not take POS n-grams into account for feature engineering, because these features seem to point to authors' non-native proficiency rather than to idiosyncratic linguistic features carried over from their L1 dialect. Taking POS n-grams into consideration would have been misleading, as it might well be possible for Peninsular Spanish speakers to have higher non-native proficiency than Mexican Spanish speakers in other data sets.

Our best performing classifier used punctuation features and scored just above chance with an accuracy of 55%. The remaining features, however, did not perform well overall, with accuracy scores ranging from 36%-50%. Interestingly, the combination of features did not perform well either.

Notably, many of these features aside from punctuation are not prevalent in every text. This feature sparsity issue leaves many texts in the test set without any of the features from the training step and leads the classifier to make poor decisions. Combining features also increases the number of features for the model to consider, but in our case, this introduced more noise into the system and accuracy did not improve.

Punctuation was a distinctive feature that was present in our texts, but it did not occur enough to be a reliable classification feature. This was the case for other features (namely, adjectives of affect, intensifiers, contracted forms, and a combination of these): from our corpus analysis we note frequency differences that are distinctive between the two classes but not common enough to make a difference in automatic classification, as can be observed in Table 5.

Feature	Precision	Recall	F-1	Accuracy
token unigrams	**0.65**	**0.65**	**0.65**	**0.65**
token bigrams	**0.65**	**0.65**	**0.65**	**0.65**
punctuation signs	**0.6**	**0.55**	**0.51**	**0.55**
adjectives of affect	0.36	0.43	0.36	0.43
intensifiers	0.36	0.36	0.36	0.36
contracted forms	0.24	0.49	0.32	0.49
combination	0.42	0.45	0.38	0.45

Table 5. Classification results for the baseline n-gram model and for models with linguistic features from qualitative analysis

While result scores were low overall, the primary focus of this section of our study is to gain insight into how the linguistic features we qualitatively analyzed from our corpus perform in a classification task. Therefore, we restricted our classification models to using these features and not other traditional features that tend to work well in NLID related tasks but are difficult to explain from a linguistic viewpoint.

To improve our classification task but maintain transparency, we considered another feature that is commonly used in authorship attribution and forensic linguistic contexts: function words. The motivation for applying function words was that they are ubiquitous enough to fix our feature sparsity issue and, unlike content words, they are not topic dependent. Therefore, we offer another round of experiments that considered function words as a feature and combined them with our main linguistic features. We provide classification results in Table 6.

Feature	Precision	Recall	F-1	Accuracy
function words	0.58	0.58	0.58	0.58
function words + punctuation	**0.69**	**0.69**	**0.69**	**0.69**
function words + contracted forms	0.58	0.58	0.58	0.58
function words + adjectives of affect	0.59	0.59	0.59	0.59
function words + intensifiers	0.53	0.53	0.53	0.53
combination	0.55	0.55	0.54	0.54

Table 6. Classification results with function words as features and function words in combination with linguistic features from qualitative analysis

The classifier that used function words alone performed better than chance and better than the previous baseline classifier. The rest of the classifier variants that combined one of the linguistic features with the function word features improved accuracy, all above chance. Punctuation was a powerful feature on its own, and with function words, it yielded the highest accuracy of all (69%). This improvement surpasses the baseline token unigram and baseline classifiers. Interestingly, combining all of these features performs worse than the other variants, and combining adjectives of affect and contracted forms with function words has little to no impact on the models. These features do not seem to correlate well together, even though they may yield different scores in isolation. Additionally, we note that all results have the same precision and recall score. This is more expected for balanced datasets. As a result, this also leads to a similar F-1 score and accuracy score.

Our best performing classifier can predict between a text produced from an L1 speaker of Peninsular Spanish and Mexican Spanish when writing in L2 English with a unique combination of features that are transparent, topic independent, generalizable, and prevalent. While using an engineering approach, and specifically a logistic regression model, is not always applicable to the forensic context because of text length and explainability issues, we show that linguistically informed features can improve an automatic approach for this task. Therefore, such a model potentially serves as an analysis tool that balances an automatic approach and its linguistic explainability, which can be helpful when tackling automatic language and dialect identification tasks. These results, however, only pertain to the scope of this specific genre and text size and need further research with data from L2 English texts written by authors with other L1 Spanish dialects. Indeed, these models would struggle with shorter texts where these features are possibly sparser.

While the accuracy seems relatively low and there is still room for improvement, these experiments are a starting point for NDID classification. There was not one feature that carried the prediction but, from among the features we analyzed, punctuation

seemed to be the most effective, suggesting that other NDID classification tasks within online genres should consider punctuation types and punctuation repetition.

Conclusion and future perspectives

We presented a first effort to identify two Spanish L1 dialects, Mexican and Peninsular, when analyzing texts written in L2 English. The mixed-approach methodology we used to tackle NDID provides a comprehensive linguistic description of features that serve to identify Mexican and Peninsular dialects of Spanish. As Kingston (2019) remarks, a frequent problem with NLI studies is that they tend to neglect the features used to classify texts; our study, on the contrary, has taken linguistic features into consideration as they are key for explainability in the forensic context.

In terms of linguistic features in our data, a careful analysis from results revealed that frequency alone is not necessarily explanatory; as an example, while *a (little) bit* is more frequent in the Peninsular Spanish corpus, it shows a wider array of pragma-linguistic functions in the Mexican Spanish corpus, mirroring what happens in speakers' L1 dialects. Thus, qualitative analyses are needed to help dissect between dialects more accurately. Another revealing conclusion from our study is that stance markers, which vary culturally, can aid the forensic linguist in discovering the L1 dialect of a language from L2 output. Additionally, use of punctuation signs in the data follows patterns found in previous studies (Bieswanger 2008; Gibson *et al.* 2018), where CMC users from different cultural backgrounds show different communicative strategies and where use of non-word tokens is culturally bound. Importantly, results showed that we can tell the difference between Spaniards' and Mexicans' L2 English output, so that it might be possible to differentiate English L2 outputs among Latin American Spanish dialects. Having said this, inter-medial and inter-modal studies are needed to further advance knowledge in NDID.

We also implemented a classifier to detect the L1 Spanish dialect of an anonymous author writing in L2 English. In the experiments we carried out on the test data, our model with function words and punctuation as features achieved accuracy of 69% in categorizing unseen Tripadvisor entries. The applicability of this method to different data requires further investigation, as we need to see if it is transferable to different genres. Furthermore, the question of how short a text can be and still allow accurate categorization still needs to be answered.

This investigation contributes to the field of authorship attribution studies in general and to NLID and NDID studies in particular by being the first of its kind to address NDID in non-contact dialects of any language. Future studies should apply the same methodology to other dialects of Spanish to see if these linguistic features apply to all dialects of Spanish in general. Finally, another issue that commands additional research is whether the same linguistic features can detect native dialect influence in L2 English (or any other L2) texts with an L1 that is not Spanish.

Notes

[1]According to its Wikipedia entry, Tripadvisor is "an American online travel company that operates a website and mobile app with user-generated content, a comparison shopping website, and offers online hotel reservations as well as bookings for transportation, lodging, travel experiences, and restaurants" (https://en.wikipedia.org/wiki/TripAdvisor).

[2]We are aware that gauging native proficiency can be complicated. In this specific case, we were looking for accuracy (that is, the ability to produce grammatically correct sentences) and grammatical range. The author who carried out the analysis is a native Spanish speaker and has also carried out L2 proficiency level assessments.

[3]Intensifiers are adverbs that magnify meaning, scaling a quality up (Ito and Tagliamonte 2003).

[4]All data for AmE frequencies in the study were obtained from the *Corpus of Contemporary American English* (COCA, Davies (2008)), while data for all BrE frequencies were obtained from the *British National Corpus* (BNC Consortium 2007).

[5]As Márquez-Reiter (2002) points out, these are quite generalized cultural perceptions: Latin America is a vast, culturally diverse geographical area.

[6]The recall score refers to the number of predicted labels divided by the number of labels in the dataset. The precision score refers to the number of predicted labels divided by the number of those predicted labels that actually belong to the label.

[7]Since we have more of an even class distribution—an even number of testing input from Mexican Spanish and Peninsular Spanish— we anticipate the accuracy score being sufficient for this study.

References

Algeo, J. (2006). *British or American English? A handbook of word and grammar patterns*. https://doi.org/10.1017/CBO9780511607240: Cambridge University Press.

Biber, D. (1987). A textual comparison of british and american writing. *American Speech*, 62(2), 99–119.

Biber, D., Gray, B. and Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly*, 45(1), 5–35.

Biber, D., Gray, B. and Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, 37, 639–668.

Bieswanger, M. (2008). 2 abbrevi8 or not 2 abbrevi8: A contrastive analysis of different space- and time-saving strategies in english and german text messages. *Texas Linguistics Forum*, 50. http://studentorgs.utexas.edu/salsa/proceedings/2006/Bieswanger.pdf.

BNC Consortium, (2007). *The British National Corpus*. Oxford Text Archive, xml ed. http://hdl.handle.net/20.500.12024/2554.

Bredel, U. (2008). *Die Interpunktion des Deutschen. Ein kompositionelles System zur Online-Steuerung des Lesens*. Tübingen: Niemeyer.

Bredel, U. (2011). *Interpunktion*. Heidelberg: Winter.

Bronshteyn, K. and Gustafson, T. (2015). The acquisition of phrasal verbs in l2 english: A literature review. *Linguistic Portfolios*, 4(1), 91–99.

Burrows, J. (2007). All the way through testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22, 27–47.

Busch, F. (2021). The interactional principle in digital punctuation. *Discourse, Context & Media*, 40, 100481.

Bäcklund, U. (1973). *The collocation of adverbs of degree in English*. Almqvist & Wiksell.

Caraker, R. (2016). Spain and the context of english language education. *Research Bulletin*, 92, 23–35.

Company Company, C. (2002). Gramaticalización y dialectología comparada: Una isoglosa sintáctico-semántica del español. *Dicenda: Cuadernos de Filología Hispánica*, 20, 39–71.

Coulthard, M., Grant, T. and Kredens, K. (2010). Forensic linguistics. In R. Wodak, B. Johnstone and P. Kerswill, Eds., *The SAGE Handbook of Sociolinguistics*. Sage Publications, 529–544.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Blackwell Publishing, 6th ed.

Davies, M. (2008). The corpus of contemporary american english (coca): 385 million words, 1990-present. https://www.english-corpora.org/coca/.

Despagne, C. (2010). The difficulties of learning english: Perceptions and attitudes in mexico. *Canadian and International Education / Education Canadienne et International*, 39(2), 55–74.

Fernández Vítores, D. (2020). El español: una lengua viva. informe. https://www.cervantes.es/imagenes/File/espanol_lengua_viva_2019.pdf.

Gibson, W., Huang, P. and Yu, Q. (2018). Emoji and communicative action: the semiotics, sequence and gestural actions of 'face covering hand. *Discourse, Context & Media*, 26, 91–99.

Goldin, G., Rabinovich, E. and Wintner, S. (2018). Native language identification with user generated content. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 3591–3601, https://aclanthology.org/D18-1000.pdf.

Grant, T. (2008). Approaching questions in forensic authorship analysis. In J. Gibbons and M. Turell, Eds., *Dimensions of Forensic Linguistics*. John Benjamins Publishing Company, 215–229.

Grant, T., Kredens, K. and Perkins, R. (2010). Identifying an author's native language phase 2 + finding and training the bilingual language expert.

Gumperz, J. (1982). *Discourse strategies*. Cambridge University Press.

Herring, S. (2012). Grammar and electronic communication. In C. Chapelle, Ed., *Encyclopedia of applied linguistics*. Wiley-Blackwell.

Honnibal, M. and Montani, I. (2015). spacy: Industrial-strength natural language processing (nlp) with python and cython). https://spacy.io.

Ito, R. and Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in english intensifiers. *Language in society*, 32(2), 257–279.

Jiang, X., Guo, Y., Geertzen, J., Alexopoulou, D., Sun, L. and Korhonen, A. (2014). Native language identification using large, longitudinal data. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, 3309–3312: European Language Resources Association (ELRA).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1(1), 7–36.

Kingston, J. (2019). *Other Language Influence of Metropolitan, Canadian, and Maghrebi French in Colloquial Written English [Unpublished Master's Thesis*. Birmingham, UK: Aston University.

Koppel, M., Schler, J. and Zigdon, K. (2005). Automatically determining an anonymous author's native language. In P. Kantor, G. Muresan, F. Roberts, D. Zeng, F.-Y. Wang, H. Chen and R. Merkle, Eds., *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics*, 209–217: Springer.

Krashen, S. (1977). The monitor model for adult second language performance. In M. Burt, H. Dulay and M. Finocchlaro, Eds., *Viewpoints on English as a second language*. Regents Publishing Company, 152–161.

Kyle, K. and Crossley, S. (2018). Measuring syntactic complexity in l2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2), 333–349.

Larsen-Freeman, D. (1978). An esl index of development. *TESOL quarterly*, 12, 439–448.

Lightbown, P. (1987). Classroom language as input to second language acquisition. In C. Pfaff, Ed., *First and second language acquisition processes*. Newbury, 169–187.

Lipski, J. (2012). Geographical and social varieties of spanish: An overview. In J. Hualde, A. Olarrea and E. O'Rourke, Eds., *The handbook of Hispanic linguistics*. https://doi.org/10.1002/9781118228098: Wiley-Blackwell, 1–26.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *TESOL quarterly*, 45(1), 36–62.

Maier, W. and Gómez-Rodríguez, C. (2014). Language variety identification in spanish tweets. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, 25–35.

Mikolov, T., Yih, W.-t. and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Márquez-Reiter, R. (2002). A contrastive study of conventional indirectness in spanish: Evidence from peninsular and uruguayan spanish. *Pragmatics*, 12(2), 135–151.

Odlin, T. (1978). Variable rules in the acquisition of english contractions. *TESOL Quarterly*, 12(4), 451–458.

Ortega, L. (2015). Syntactic complexity in l2 writing: Progress and expansion. *Journal of Second Language Writing*, 29, 82–94.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.

Perkins, R. (2015). Native language identification (nlid) for forensic authorship analysis of weblogs. In M. Dawson and O. Marwan, Eds., *New threats and countermeasures in digital crime and cyber terrorism*. https://doi.org/10.4018/978-1-4666-8345-7.ch012: IGI Global, 213–234.

Perkins, R. and Grant, T. (2013). Forensic linguistics. In J. Siegel, P. Saukko and M. Houck, Eds., *Encyclopedia of Forensic Sciences*. Elsevier, 174–177.

Precht, K. (2000). *Patterns of stance in English*. Northern Arizona University.

Precht, K. (2003a). Great versus lovely: Stance differences in american and british english. In P. Leystina and C. Meyer, Eds., *Corpus Analysis: Language structure and language use*. Brill Rodopi, 133–151.

Precht, K. (2003b). Stance moods in spoken english: Evidentiality and affect in british and american conversation. *Text & Talk*, 23(2), 239–257.

Rangel, F., Franco-Salvador, M. and Rosso, P. (2016). A low dimensionality representation for language variety identification. In *International Conference on Intelligent Text Processing and Computational Linguistics*, 156–169, Cham: Springer.

Reynoso Noverón, J. (2001). La pragmática como evidencia en el contacto español-lenguas indígenas. el diminutivo en el español actual. In C. Matute and A. Palacios, Eds., *El indigenismo americano (II). Actas de las Segundas Jornadas Internacionales sobre Indigenismo Americano*. Universitat de València, Facultat de Filología, 213–222.

Sadat, F., Kazemi, F. and Farzindar, A. (2014). Automatic identification of arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, 22–27: Association for Computational Linguistics and Dublin City University.

Samar, R. (2003). Aux-contraction in second language speech: A variationist analysis. *Cahiers Linguistiques d'Ottawa*, 31.

Snider, N. and Arnon, I. (2012). A unified lexicon and grammar? compositional and non-compositional phrases in the lexicon. In D. Divjak and S. Gries, Eds., *Frequency effects in language representation*. https://doi.org/10.1515/9783110274073.127: De Gruyter-Mouton, 127–163.

Terkourafi, M. (2011). From politeness1 to politeness2: tracking norms of im/politeness across time and space. *Journal of Politeness Research*, 7(2), 159–185.

Tetreault, J., Blanchard, D. and Cahill, A. (2013). A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, 48–57, The Association for Computational Linguistics. https://aclanthology.org/W13-1700.pdf.

Thomason, S. (2001). *Language Contact: An Introduction*. Georgetown University Press.

Travis, C. (2004). The ethnopragmatics of the diminutive in conversational colombian spanish. *Intercultural Pragmatics*, 1(2), 249–274.

Tyler, A. and Evans, V. (2001). Reconsidering prepositional polysemy networks: The case of over. *Language*, 77(4), 724–765.

Wierzbicka, A. (1984). Diminutives and depreciatives: Semantic representation for derivational categories. *Quaderni di semantica*, 5(1), 123–130.

Wierzbicka, A. (1992). *Semantics, culture, and cognition: Universal human concepts in culture-specific configurations*. Oxford University Press.

Wong, S. and Dras, M. (2011). Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1600–1610.

Yaeger-Dror, M. (1997). Contraction of negatives as evidence of variation in register specific interactive rules. *Language Variation and Change*, 9, 1–36.

Young, J. (2015). *Exploring Japanese learners' perception, production, and beliefs concerning spoken English contractions*. University of Illinois at Urbana-Champaign.] ProQuest Dissertations Publishing.

Zampieri, M. and Gebrekidan-Gebre, B. (2012). Automatic identification of language varieties: The case of portuguese. In *KONVENS2012-The 11th Conference on Natural Language Processing*, 233–237: Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI).