Native Language Identification in Texts: A Survey

Dhiman Goswami[1], Sharanya Thilagan[1], Kai North[1], Shervin Malmasi[2], Marcos Zampieri[1]
[1]George Mason University, Fairfax, VA, USA
[2]Amazon.com, Inc. Seattle, WA, USA
dgoswam@gmu.edu

Abstract

We present the first comprehensive survey of Native Language Identification (NLI) applied to texts. NLI is the task of automatically identifying an author's native language (L1) based on their second language (L2) production. NLI is an important task with practical applications in second language teaching and NLP. The task has been widely studied for both text and speech, particularly for L2 English due to the availability of suitable corpora. Speech-based NLI relies heavily on accent modeled by pronunciation patterns and prosodic cues while text-based NLI relies primarily on modeling spelling errors and grammatical patterns that reveal properties of an individuals' L1 influencing L2 production. We survey over one hundred papers on the topic including the papers associated with the NLI and INLI shared tasks. We describe several text representations and computational techniques used in text-based NLI. Finally, we present a comprehensive account of publicly available datasets used for the task thus far.

1 Introduction

Native and non-native speakers of any language are well-equipped to recognize foreign accents in non-native speech (Major, 2007). Pronunciation, stress, and prosodic patterns that diverge from those used by native speakers (L1) are widely used to automatically identify a second language (L2) speaker's mother tongue, a task known as speech-based NLI (Krishna et al., 2019). The same way we hear foreign accents, it is also possible to identify non-native linguistic patterns, mostly related to word choices, syntax, and spelling, in texts written by non-native authors. Computational models can therefore be trained on texts belonging to non-native writers to learn common properties of a given L1 with the goal of determining the writer's mother tongue, a task known as text-based NLI (Malmasi, 2016).

The fundamental assumption behind NLI is that the mother tongue influences Second Language Acquisition (SLA) and production; the latter is known as cross-linguistic influence (Krashen, 1981; Jarvis and Pavlenko, 2008; Ellis, 2015). Language transfer may occur in SLA (Gass, 1988) and it can have negative or positive influence in acquisition. Negative transfer occurs when differences between the two languages structures lead to systematic errors in the learning of the L2. This leads to a process called fossilization when the use of particular incorrect patterns becomes a habit thus posing communication problems to the L2 speaker (Han and Odlin, 2005). Positive transfer occurs when areas of similarity (e.g., similar words) between the two languages facilitate learning. Due to such positive and negative transfer from the L1 to the L2, it becomes possible to train computational models to recognize patterns shared by speakers of the same L1 when speaking or writing in an L2.

There are several reasons to use models to study non-native texts and perform text-based NLI. Firstly, computational models can be used to investigate the influence of native language in SLA (Jarvis and Crossley, 2012; Jarvis et al., 2019). This can be used to better understand language transfer (Liu et al., 2022) and the findings can be applied to the development of L2 teaching materials and Computer-aided Language Learning (CALL) software. Secondly, NLI can improve NLP systems for processing texts written by non-native speakers (Rabinovich et al., 2016) in applications such as machine translation (Anastasopoulos et al., 2019). NLI is also relevant to forensic linguistics which often entails authorship attribution of disputed documents (Malmasi, 2016) or detecting plagiarism (Malmasi et al., 2017a).

Despite the importance of this task, to the best of our knowledge, no comprehensive survey of text-based NLI exists. This paper provides the first comprehensive survey on the topic. We survey over

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3149–3160

100 papers on the topic published primarily at the ACL Anthology, but also at other repositories such as the ACM Digital Library. As evidenced in this survey, most work in NLI has relied on feature engineering and traditional machine learning (ML) classifiers (e.g., SVMs) (Jarvis et al., 2013; Malmasi and Dras, 2014b; Vajjala and Banerjee, 2017). Recent advances in deep learning, and the introduction of Large Language Models (LLMs) have brought renewed interest in NLP, opening exciting new avenues for new and existing tasks. We believe that by reviewing prior work, this survey can identify knowledge gaps and open research questions that can help shape future research.

The remainder of this paper is organized as follows: Section 2 presents an overview of text-based NLI. Section 3 provides a brief description of available NLI datasets. Section 4 discusses the two most well-known NLI shared tasks. Section 5 describes features used in NLI while Section 6 discusses the different models that have been used to approach NLI. Lastly, Section 7 presents concluding remarks and discusses avenues for future work.

2 NLI Task Overview

Text-based NLI is typically modeled as a supervised multi-class classification task evaluated using widely-used metrics in text classification such as accuracy, precision, recall, and F-score (Malmasi, 2016). Given a collection of texts in a particular L2 written by a set of speakers from n different L1s, NLI systems are trained to assign a class label (L1) to each document in the text collection. Figure 1 illustrates the NLI task with an English L2 and four different L1s.

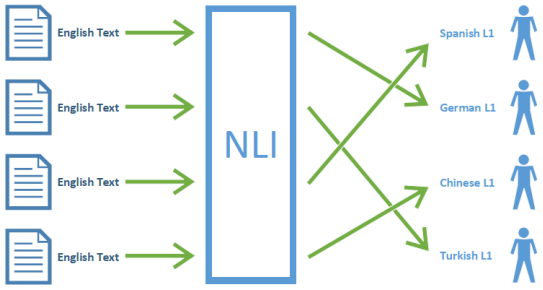

Figure 1: The general concept of an NLI system as depicted by Malmasi (2016).

As discussed in this survey, NLI systems rely on a variety of features such as words, characters, morphosyntactic information, dependency relations, and spelling errors. While word-based features (e.g., word n-grams) have shown to deliver high performance in most text classification tasks, *topic bias* is a significant confounder. Topic bias occurs when thematic references in a text reveal the true L1 label, for example, English texts written by L1 speakers of Japanese are more likely to talk about Tokyo or Kyoto than texts from Chinese or Italian L1 speakers. Therefore, to decrease topic influence, several NLI studies have incorporated linguistically-motivated features that can capture morphological and syntactic variation between different L1s (Malmasi and Dras, 2014b,a; Malmasi et al., 2018, 2015a; del Río, 2020).

3 Datasets

We present a comprehensive account of all available datasets created for and/or used in NLI research to date. The list is presented in Table 1 and summarized below.

Arabic The Arabic Learner Corpus (ALC) includes 329 in-class essays written by speakers of seven different L1's who studied Arabic as L2 in Saudi Arabia.

Chinese The Chinese Learner Corpus (CLC) features 3,216 essays of 600 token on an average and written in Chinese by university students from 11 countries, representing diverse proficiency levels.

Czech CzeSL-SGT is a subset of the publicly available Czech as a Second Language with Spelling, Grammar, and Tags corpus, comprising Czech 8,617 essays written by 54 different L1s. The used dataset includes 47% Non-Indo European, 29% Slavic and 24% Indo European L1's after excluding texts of unknown language group.

English This is the most well-studied L2 in NLI. There are ten datasets presented in Table 1. The most popular is the TOEFL11 dataset that contains 12,100 essays of eight different prompts and three score levels written by speakers of eleven L1s.

Finnish The Corpus of Advanced Learner Finnish (LAS2) comprises 204 Finnish L2 writings from speakers seven different L1s collected as part of a project at the University of Turku.

German The FALKO (fehlerannotierten Lernerkorpus) corpus is the largest publicly available selection of German learner texts. It contains 221 essays and text summaries written by eight different non-native speakers of German.

L2	Dataset	L1s	Instances	Reference
Arabic	ALC	Chinese, Urdu, Malay, French, Fulani, English, Yoruba	329	(Malmasi and Dras, 2014a)
Chinese	CLC	Filipino, Indonesian, Thai, Laotian, Burmese, Korean, Khmer, Vietnamese, Japanese, Spanish, Mongolian	3,216	(Malmasi and Dras, 2014b)
Czech	CzeSL-SGT	54 different L1's (5 most frequent languages - Chinese, Russian, Ukrainian, Korean, English - 3,715 instances)	8,617	(Tydlitátová, 2016)
English	CLC FCE	Catalan, Chinese, Dutch, French, German, Greek, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, Turkish	1,244	(Malmasi, 2016)
	Estival	English, Spanish, Arabic	9,836	(Estival et al., 2007)
	ESL	Chinese, Czech, Italian, Russian, Spanish	4,185	(Rozovskaya and Roth, 2011)
	ICLE v2.0	Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish, Tswana	6,085	(Granger et al., 2009)
	INLI 2017	Bengali, Hindi, Kannada, Malayalam, Tamil, Telugu	10,647	(Anand Kumar et al., 2017)
	ICNALE	ENL (US, UK, Australia etc.), ESL (Hong Kong, Pakistan), EFL (China, Indonesia, Japan, Korea, Taiwan, Thailand)	4,428	(Ishikawa, 2011)
	Lang-8	65 different native languages included, with 14 of those languages having 1000 entries (Japanese, Chinese (Mandarin and Cantonese), Korean, Russian, Spanish, French, German, Polish, Italian, Vietnamese, Indonesian, Arabic, Portuguese, Thai)	154,702	(Brooke and Hirst, 2011)
	OGI-TS	English, Farsi, French, German, Hindi, Korean, Japanese, Mandarin, Spanish, Tamil, Vietnamese	1,236	(Zissman, 1996)
	Reddit-L2	English, German, Dutch, French, Polish, Romanian, Finish, Swedish, Spanish, Greek, Portuguese, Estonian, Czech, Italian, Russian, Turkish, Bulgarian, Croatian, Norwegian, Hungarian, Lithuanian, Slovenian and Serbian	200 million	(Rabinovich et al., 2018)
	TOEFL11	Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish	12,100	(Blanchard et al., 2013)
Finnish	LAS2	Russian, Japanese, Lithuanian, Czech, German, Hungarian, Polish, Komi, English	204	(Malmasi and Dras, 2014c)
German	FALKO	Chinese, Danish, English, French, Polish, Russian, Turkish, Uzbek	221	(Malmasi and Dras, 2017a)
Italian	VALICO	Albanian, Chinese, Czech, English, French, German, Hindi, Japanese, Polish, Portuguese, Romanian, Russian, Serbian, Spanish	2,531	(Malmasi and Dras, 2017a)
Norwegian	ASK	German, Dutch, English, Spanish, Russian, Polish, Bosnian-Croatian-Serbian, Albanian, Vietnamese, Somali	2,158	(Malmasi et al., 2015a)
Portuguese	NLI-PT	Chinese, English, Spanish, German, Russian, French, Japanese, Italian, Dutch, Tetum, Arabic, Polish, Korean, Romanian, Swedish	1,868	(del Río et al., 2018)
Russian	RLC	Chinese, Danish, English, Estonian, Finnish, French, Deutsch, Italian, Japanese, Kazakh, Korean, Norwegian, Serbian, Swedish, Thai	7,831	(Remnev, 2019)
Spanish	ARU	English, French, German, Greek, Italian, Japanese	206	(Malmasi and Dras, 2017a)
Turkish	TLC	Arabic, Albanian, Azeri Turkish, Farsi, Afghani	284	(Uluslu, 2023)

Table 1: A list of LI datasets with L2, L1s, number of instances, and a reference to the paper describing the resource.

Italian VALICO (Varietà di Apprendimento della Lingua Italiana Corpus Online) includes approximately one million tokens of learner Italian writing from a wide range of L1s along with the associated metadata. A subset containing 2,531 texts from 14 different L1s is used for NLI.

Norwegian The ASK Corpus is a learner corpus composed of the 2158 essays of learners of Norwegian who are native speaker of 10 different L1s. These texts are essays written as part of a test of Norwegian as a second language.

Portuguese NLI-PT was collected from three different learner corpora of Portuguese: (i) COPLE2; (ii) Leiria corpus, and (iii) PEAPL2, containing 148 topics including written exercises on lessons, official Portuguese proficiency test and different stimuli from learners of Portuguese with different proficiency levels and fifteen L1s.

Russian The Russian Learner Corpus (RLC) is a collection of 7831 - both academic and non-academic texts produced by speakers of fifteen L1s.

Spanish The Anglia Ruskin University (ARU) Spanish learner corpus contains 206 texts written by speakers of six L1s who are mainly undergraduate students of the university and some ERASMUS students. The texts that were produced by students either as course work or as part of exams.

Turkish The Turkish Learner Corpus (TLC) covers 284 essays written by L1 speakers of Afghan, Albanian, Arabic, Azeri Turkish and Farsi.

4 Shared Tasks

Shared tasks are competitions in which multiple participating teams develop systems to address a task using the same benchmark dataset typically provided by the shared task organizers. The goal is to foster research on a topic while encouraging wide collaboration (Escartín et al., 2017).

Shared tasks provide important benchmarks and datasets to the community. Several related shared tasks on language and dialect identification have been organized (Zubiaga et al., 2016; Gaman et al., 2020; Aepli et al., 2022), but only two shared tasks on text-based NLI have been organized thus far. The two editions of the NLI shared tasks co-located with the workshop on Innovative Use of NLP for Building Educational Applications (BEA) (Tetreault et al., 2013; Malmasi et al., 2017b) included English texts written by speakers of 11 L1s. The Indian Native Language Identification (INLI) shared task at the Forum for Information Retrieval Evaluation (FIRE) 2017 (Anand Kumar et al., 2017) and 2018 (Kumar et al., 2018) featured an English dataset with comments from social media written by speakers of multiple L1s spoken in India such as Bengali, Hindi, Kannada, Malayalam, Tamil, and Telugu. We describe the two editions of each of these shared tasks in detail in this section.

Native Language Identification 2013 (NLI-2013) NLI-2013 (Tetreault et al., 2013) comprised three sub-tasks using the TOEFL11 dataset containing English texts written by speakers of 11 L1s: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. The submitted systems were evaluated on the TOEFL11 test set while training data varied according to the sub-task. In (1) closed-training, participants could use only the TOEFL11 training set and no additional external data; in (2) open-training-1: participants could use any external dataset for training but not TOEFL11; finally in (3) open-training-2 participants could used the TOEFL11 training set in combination with any other dataset. Each participant could submit up to five systems for each of the three sub-tasks. This allowed teams to experiment with different variations of their core system. The best performing teams approached the task using classic ML systems such as SVMs, Logistic Regression, String Kernels, Local Rank Distance, and Maximum Entropy. The top systems were trained on features such as word and Part-of-Speech (POS) n-grams (Goutte et al., 2013; Gebre et al., 2013; Tsvetkov et al., 2013; Jarvis et al., 2013).

Native Language Identification 2017 (NLI-2017) NLI-2017 (Malmasi et al., 2017b) featured text and speech. The organizers provided participants with a corpus including written essays and transcriptions of spoken responses from TOEFL proficiency test takers. The same 11 L1s as NLI 2013 were included in NLI 2017. The task featured three tracks: essay-only, speech-only, and fusion. In the essay-only track, participants predicted the candidate's L1 based only on essays. In the speech-only track, participants predicted the L1 based on a transcription of a 45-second spoken response. The raw audio for the spoken responses was not distributed, but i-vectors, low-dimensional representations of acoustic measurements (Dehak et al., 2011; Martinez et al., 2011), were provided. Finally, in the fusion track, participants predicted the L1 using a combination of written essays and spoken responses. Both open and closed competitions were held, allowing participants to use additional NLI training data in the open competition but limiting the use of external data in the closed competition. Systems featuring ensemble of multiple traditional ML classifiers (e.g. SVMs) trained on lexical and syntatic features were the most effective approaches in all tracks (Goutte and Léger, 2017; Li and Zou, 2017).

Indian Native Language Identification 2017 (INLI-2017) INLI-2017 (Anand Kumar et al., 2017) provided participants with a corpus containing over 10,600 Facebook comments in English written by native speakers of Bengali, Hindi, Kannada, Malayalam, Tamil, and Telugu. The comments were retrieved from local newspaper pages. Instances containing code-mixing, a wide-spread linguistic phenomenon in the region, have been excluded. Most high-performing approaches in this competition used TF-IDF vectors with SVMs (Kosmajac and Keselj, 2017; Lakshmi and Shambhavi, 2017). Deep learning approaches have not obtained competitive performance in INLI-2017 displaying inferior performance compared to the SVM baseline system released by the organizers.

Indian Native Language Identification 2018 (INLI-2018) INLI-2018 (Kumar et al., 2018) was the second iteration of the INLI shared task. It features English Facebook comments written by speakers of the same six L1s included in the previous edition. The organizers provided participants with the same training data as the 2017 edition and

two test sets. Test set 1 was the same test set from the 2017 edition while test set 2 was a novel test set compiled one year after the training data. The best performing systems in NLI-2018 used TF-IF vectors as features along with neural networks and ensemble approaches combining predictions from both traditional ML and deep learning classifiers (Ajees and Idicula, 2018; Nayel and Shashirekha, 2018; Markov and Sidorov, 2018). The use of neural network architectures with word embeddings has not been explored in this shared task.

5 Text Representation and Features

Most approaches to NLI have relied on statistical ML classifiers. These classifiers take a variety of features as input such as characters and, most commonly, words extracted from the training corpora often as part of n-gram models. In addition to character and word-based features, POS tags are widely used in NLI along with several additional lexical and syntactic features. These different feature classes have been shown to provide complementary information (Malmasi and Cahill, 2015). In the following paragraphs we define each feature and provide a list of studies that use each group of feature in Table 2.

Lexical Features Lexical features used in NLI include word frequencies, word type frequencies, spelling errors, number of phonemes and syllables, and more. These features provide insight into the linguistic fingerprint of an individual and can therefore be used to identify their native language. Swan and Smith (2001) and Gebre et al. (2013) used spelling and grammatical errors to predict learners' L1 influences for NLI. del Río (2020) and Malmasi and Dras (2014c) used frequencies of letters, phonemes, syllables, morphemes and suffixes respectively for Portuguese and Finnish NLI. Wong and Dras (2009) and Malmasi and Dras (2017a) explored the use of function words for NLI, including the use of stopword lists for multilingual NLI.

N-grams N-grams are a sequence of n items in texts, most often letters or words, and are subsequently defined at either the character-level or word-level. Wong and Dras (2009) utilized word-level n-grams and character bigrams and trigrams, and the 100 most frequent unigrams found within their dataset. Gebre et al. (2013) also explored word-level unigrams and bigrams both separately and in combination. Tydlitátová (2016) employed character n-grams extracted from individual words for n = 1, 2, 3 for Czech. N-grams therefore have been frequently used throughout NLI research to identify character or word combinations that are indicative of an individual's native language. N-grams have also been used to model sequences of POS tags as described in this section.

POS tags Part-of-Speech (POS) tags are used to denote a word's grammatical function, such as noun, verb, adjective, etc. POS tags are assigned at the word-level most commonly by an automatic POS tagger. Several studies have used POS tag frequencies to help differentiate between speakers of various L1s. Malmasi et al. (2015a) used Mixed POS function word n-grams (n = 1-3) influenced by Wong et al. (2012) for NLI of Norwegian texts. Malmasi et al. (2018) and del Río (2020) used fine grained POS tags as topic independent features of NLI including verbs, nouns and adjectives for Portuguese. Mechti et al. (2016) and Mechti et al. (2020) used POS n-gram (n = 1..3) and applied Alkhalil Arabic-specific morphological analyser (Boudlal et al., 2010) to tags word for Arabic NLI.

TF-IDF Weighting While the aforementioned features can be represented by binary or normalized feature vectors, Term Frequency-Inverse Document Frequency (TF-IDF) weighting takes into account feature frequency across the entire training corpus, helping to identify rare features that might be highly discriminative. TF-IDF vectors have been frequently used for information retrieval and various NLP tasks such as text categorization and authorship identification. Their ability to characterize a given text along with their predictive capabilities has made their use popular in NLI, often leading to improved performance (Gebre et al., 2013; Zampieri et al., 2017).

Syntactic Features Syntactic transfer from L1 to L2 has been widely studied in various language pairs (Liu et al., 2022). Syntactic features for NLI range from the use of Context Free Grammar (CFG) structures to the frequency of noun or verb phrases within a text. Mechti et al. (2020) explored several syntactic features including the use of the CFG production rule. Wong and Dras (2011) proposed a model treating parse tree horizontal slices as CFG production rule sets, using them as binary features. In addition, several studies have investigated topic-independent features alongside CFG rules to capture relationships useful for NLI, includ-

Dataset	Char-N-grams	Word-N-grams	POS tags	TF-IDF	Lexical	Syntactic	Reference
ICLE	✓	-	✓	-	✓	-	(Koppel et al., 2005)
ICLE	✓	✓	-	-	-	-	(Wong and Dras, 2009)
ICLE	-	-	-	-	-	✓	(Swanson and Charniak, 2012)
ICLE	-	-	-	-	-	✓	(Tetreault et al., 2012)
ICLE	-	✓	✓	-	-	-	(Wong et al., 2012)
TOEFL11	-	✓	-	-	-	-	(Wu et al., 2013)
TOEFL11	✓	✓	-	✓	✓	-	(Lynum, 2013)
TOEFL11	-	-	-	-	-	✓	(Mizumoto et al., 2013)
TOEFL11	✓	✓	✓	✓	-	✓	(Gebre et al., 2013)
TOEFL11	✓	✓	✓	-	-	-	(Jarvis et al., 2013)
Multiple	✓	✓	✓	-	✓	✓	(Bykh et al., 2013)
TOEFL11	✓	✓	✓	-	-	✓	(Goutte et al., 2013)
TOEFL11	✓	✓	-	-	-	-	(Henderson et al., 2013)
TOEFL11	-	-	✓	-	✓	-	(Tsvetkov et al., 2013)
ALC	-	-	✓	-	-	✓	(Malmasi and Dras, 2014a)
CLC	-	-	✓	-	-	✓	(Malmasi and Dras, 2014b)
LAS2	-	✓	✓	-	-	-	(Malmasi and Dras, 2014c)
ASK	-	-	-	-	-	✓	(Malmasi et al., 2015a)
TOEFL11	✓	✓	✓	-	-	✓	(Malmasi et al., 2015b)
ALC	-	-	-	-	-	✓	(Mechti et al., 2016)
CzeSL-SGT	-	-	✓	-	-	-	(Tydlitátová, 2016)
TOEFL11	-	✓	-	-	✓	-	(Mohammadi et al., 2017)
Multiple	-	-	✓	-	-	-	(Malmasi and Dras, 2017a)
TOEFL11	✓	✓	-	✓	-	-	(Zampieri et al., 2017)
INLI	-	-	-	✓	-	-	(Nayel and Shashirekha, 2017)
INLI	-	-	-	✓	-	-	(Bharathi et al., 2017)
Multiple	✓	✓	✓	-	-	-	(Ircing et al., 2017)
TOEFL11	✓	✓	✓	-	-	-	(Vajjala and Banerjee, 2017)
Multiple	✓	✓	✓	-	✓	-	(Markov et al., 2017)
Multiple	-	✓	✓	-	✓	-	(Kulmizev et al., 2017)
Multiple	✓	✓	-	-	-	-	(Rama and Çöltekin, 2017)
Multiple	✓	✓	-	-	✓	✓	(Li and Zou, 2017)
Multiple	✓	✓	✓	-	-	-	(Goutte and Léger, 2017)
Multiple	✓	✓	-	-	-	-	(Oh et al., 2017)
Multiple	✓	-	-	-	✓	✓	(Bjerva et al., 2017)
INLI	-	-	-	-	✓	✓	(Anand Kumar et al., 2017)
INLI	-	-	-	✓	✓	-	(Jain et al., 2017)
INLI	✓	✓	✓	-	-	-	(Markov and Sidorov, 2018)
INLI	-	-	-	✓	✓	-	(Gupta, 2018)
INLI	-	-	-	✓	-	-	(Mondal et al., 2018)
INLI	-	-	-	✓	✓	-	(Thenmozhi et al., 2018)
RLC	✓	✓	-	✓	-	-	(Remnev, 2019)
NLI-PT	-	✓	-	-	-	-	(del Río, 2020)
ALC	-	-	-	✓	-	✓	(Mechti et al., 2020)
TLC	✓	✓	✓	-	✓	-	(Uluslu, 2023)

Table 2: Primary features used in ML and feature engineering approaches to NLI along with references.

ing for Chinese (Malmasi and Dras, 2014b), Arabic (Malmasi and Dras, 2014a), Finnish (Malmasi and Dras, 2014c), Norwegian (Malmasi et al., 2015a) and Portuguese NLI (Malmasi et al., 2018). Swanson and Charniak (2012) and Tetreault et al. (2012) alternatively explored the use of Tree Substitution Grammars (TSGs).

6 Computational Models

A variety of traditional ML classifiers trained on the features described in Section 5 have been explored in NLI, most notably SVMs and Logistic Regression. The use of multiple base classifiers as part of ensemble or meta-classifiers have also been a popular approach for NLI yielding competitive performance in NLI-2017. Finally, various papers have approached NLI using deep learning architectures. Unlike in other NLP tasks, the performance of these approaches is, however, not always clearly superior to the performance of traditional ML classifiers.

Support Vector Machines Linear SVM classifiers are the most widely used in NLI due to their high performance and ability to deal with large and

sparse feature spaces (Jarvis et al., 2013; Malmasi and Dras, 2014b, 2017a). SVMs have been used in early work (Koppel et al., 2005) and in the highest performing entries of NLI-2013 (Gebre et al., 2013; Jarvis et al., 2013; Bykh et al., 2013; Goutte et al., 2013; Henderson et al., 2013), INLI-2017 (Markov et al., 2017; Kulmizev et al., 2017; Rama and Çöltekin, 2017; Li and Zou, 2017), and INLI-2018 (Mondal et al., 2018; Markov and Sidorov, 2018). SVM implementations available at LIBLINEAR (Fan et al., 2008) and scikit-learn (Pedregosa et al., 2011) have been widely used (Malmasi and Dras, 2014a,c; Malmasi et al., 2015a; Uluslu, 2023; Remnev, 2019; Malmasi and Dras, 2017a). Some studies (Mechti et al., 2016, 2020) have used the SVM implementation available in the LIBSVM package (Chang and Lin, 2011) whereas Tydlitátová (2016) applied SVMs with linear and polynomial kernel functions to NLI.

Logistic Regression Logistic Regression is another popular model used in NLI. Di Nuovo et al. (2020) performed guided logistic regression among four L1s in an Italian L2 dataset. In NLI-2013 various teams used Logistic Regression (Tsvetkov et al., 2013; Popescu and Ionescu, 2013). In NLI-2017 Ircing et al. (2017) used a Logistic Regression meta-classifier trained on words characters and POS base models and Vajjala and Banerjee (2017) used Logistic Regression model trained on word n-grams (1-3) for essays and transcripts of speech. Finally, Logistic Regression has also been used in INLI-2018 (Gupta, 2018).

Ensembles and Meta-classifiers An ensemble classifier combines the predictions of multiple individual classifiers, referred to as base classifiers or weak learners, to create more accurate predictive models. This combination is done by aggregating the predictions using methods such as voting. The objective of ensemble classification is to use the diversity among these base classifiers to achieve better prediction than any individual classifier. Researchers have employed ensemble methods to integrate a variety of models and features in NLI (Zampieri et al., 2017; Malmasi et al., 2018). In the NLI shared task 2017, for example, several teams used ensemble methods with competitive performance taking advantage of the multimodal nature of the dataset that contained both text and speech Goutte and Léger (2017); Oh et al. (2017); Bjerva et al. (2017); Ircing et al. (2017); Anand Kumar et al. (2017); Thenmozhi et al. (2017); Jain et al. (2017). Malmasi et al. (2015b) combined the predictions from all NLI Shared Task 2013 systems in a voting ensemble, showing that this approach can achieve very high accuracy results. Similar to voting ensembles, meta-classifiers also leverage the prediction of multiple base classifiers. By training a model that uses the base classifier predictions as inputs, such systems are better able to combine this information by learning patterns in base model predictions and thereby yielding more accurate predictions compared to any individual base classifier. Malmasi and Dras (2017b) presented a thorough examination of meta-classification models for NLI, achieving state-of-the-art results on three datasets from different languages. In another study, Malmasi and Dras (2018) presented a system with 200 meta-classifiers merged in a bagging ensemble, where all models are trained on different subsets of the base classifiers.

Deep Learning Deep learning approaches such as deep neural networks (DNN) (Oh et al., 2017) have also been applied to NLI with various levels of success. Examples of architectures used include Gated Recurrent Unit (GRU) (Bhargava et al., 2017), Rectified Linear Unit (ReLU) (Thenmozhi et al., 2017), Convolutional Neural Network (CNN) (Ajees and Idicula, 2018), and Long Short-Term Memory (LSTM) (Mundotiya et al., 2018). These architectures typically take an embedding representation of hundreds of dimensions instead of the n-gram models and linguistically-inspired features described in Section 5. Transformer architectures such as BERT (Devlin et al., 2019) had great impact in NLP (Rogers et al., 2021) paving the way for the latest generation of pre-trained LLMs. However, they have not had the same impact on NLI with only a few studies proposing the use of transformers for the task (Steinbakken and Gambäck, 2020; Vian, 2023). The study by Lotfi et al. (2020) showed that the performance of LSTMs and BERT on TOEFL11 and ICLE was much lower than the performance of SVMs and that of GPT-2 as described next.

LLMs and Recent Advances Recently proposed LLMs (e.g., GPT, Mistral, Llama-2) have displayed state-of-the-art performance on various NLP tasks (Minaee et al., 2024). One of the first papers to explore LLMs in NLI is the work by Lotfi et al. (2020). They have fine-tuned a GPT-2 model showing performance slightly superior to SVMs on both the TOEFL11 and ICLE corpora. More recently, a

couple studies (Uluslu and Schneider, 2022; Zhang and Salle, 2023) evaluated the performance of LLMs on the TOEFL11 dataset, including GPT-2, GPT-3.5 (Brown et al., 2020) and GPT-4 (Achiam et al., 2023). Zhang and Salle (2023) demonstrated that GPT-4 achieved state-of-the-art performance of 91.7% accuracy on this dataset using zero-shot prompting without any task-specific fine-tuning, surpassing all previously proposed ensemble approaches. Zhang and Salle (2023) further explored the interpretability of models by prompting GPT-4 to provide explanations for its predictions including linguistic features such as spelling errors, syntactic patterns, and translated phrases as cues. This study highlights the potential of generative models in achieving strong performance in NLI while also enabling explainable predictions. However, as this analysis has been done only on a single English L2 dataset, the impact of LLMs in NLI is still unclear, particularly for L2s other than English where LLMs support is still very limited.

7 Conclusion and Future Directions

This paper presented the first comprehensive survey on text-based NLI. We surveyed over 100 papers and discussed features, models, and datasets used for NLI. We collected information about all available datasets created for or used in NLI research to date. Moreover, we also discussed the NLI shared tasks organized in 2013 and 2017 and the INLI shared tasks organized in 2017 and 2018.

This survey shows that apart from a few studies (Lotfi et al., 2020; Uluslu and Schneider, 2022; Zhang and Salle, 2023) the bulk of NLI research has focused on approaches that rely on feature engineering and traditional ML classifiers, most notably SVMs. Traditional ML approaches have shown to deliver competitive performance in NLI shared tasks and other available NLI datasets. This indicates that text-based NLI is essentially a pattern-matching task that requires little to no semantic understanding of language. These findings are similar to what we observe in related language identification tasks such as dialect and language variety identification where classical ML approaches often outperform deep learning approaches (Jauhiainen et al., 2019; Zampieri et al., 2020, 2024). The similarity between NLI and other language identification has been noted in previous work as discussed in Malmasi et al. (2017b).

Amidst exciting developments in AI and NLP such as the introduction of pre-trained LLMs, we believe that this survey is an important information source for future research in text-based NLI. We expect it to be a useful resource for new and well-established researchers alike. We conclude by discussing future directions and open challenges in text-based NLI including the use of LLMs, explainability, and low-resource scenarios.

Low-resource domains As described in Section 3, most corpora used in NLI are collected within educational settings (e.g., TOEFL exams, classroom activities). L2 production, however, is not restricted to education. Texts on social media, blogs, online reviews, etc. are also written by L2 speakers and can be used in NLI (Anand Kumar et al., 2017; Rabinovich et al., 2018). More research should be carry out to produce suitable datasets that allow us to investigate L2 production in non-educational domains. This is important as prior work has shown that NLI models may not generalize across datasets (Malmasi and Dras, 2015).

Low-resource languages There is scarcity of corpora with data from L2 speakers of low-resource languages particularly from Africa, Southeast Asia, and Oceania. This hinders the development of language technology for low-resource languages, an issue addressed by initiatives such as Masakhane (Orife et al., 2020) and No Language Left Behind (NLLB) (Costa-jussà et al., 2022). The datasets presented in Section 3 mostly contain widely-spoken languages such as Arabic, Chinese, English Spanish, and Portuguese and a few mid-resource languages with millions of speakers such as Finnish and Norwegian. There is ample scope for NLI research in low-resource languages and it must start by data curation.

Large Language Models With the exception of very few studies (Lotfi et al., 2020; Zhang and Salle, 2023), LLMs have not been substantially explored for NLI. With recent advances in LLMs, we believe that more attention should be devoted to LLM-based approaches in NLI. We see task fine-tuning and prompt engineering, potentially using prompts with linguistic features indicative of L1, as two promising avenues to improving NLI performance with LLMs. Results described in Lotfi et al. (2020) indicate that task fine-tuning is a promising approach for NLI. Finally, LLMs can also be incorporated into ensemble systems which perform well with classic ML in NLI.

Explainability Explainability of models is an important area that can reveal insights about both the data and the models. The clear majority of studies discussed in this survey, however, focus on exploring which features and/or computational models work best for text-based NLI. Only a few studies have explored data-driven methods to gain insights on language transfer and/or on computational models of L2 production (Liu et al., 2022; Berti et al., 2023). With the availability of LLMs, we see ample room for more studies on model explainability in text-based NLI.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the vardial evaluation campaign 2022. In *Proceedings of VarDial*.

AP Ajees and Sumam Mary Idicula. 2018. Inli@ fire-2018: A native language identification system using convolutional neural networks. In *Proceedings of FIRE*.

M Anand Kumar, HB Barathi Ganesh, Shivkaran Singh, KP Soman, and Paolo Rosso. 2017. Overview of the inli pan at fire-2017 track on indian native language identification. In *Proceedings of CLEF*.

Antonios Anastasopoulos, Alison Lui, Toan Q Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. In *Proceedings of NAACL*.

Barbara Berti, Andrea Esuli, and Fabrizio Sebastiani. 2023. Unravelling interlanguage facts via explainable machine learning. *Digital Scholarship in the Humanities*, 38(3):953–977.

B Bharathi, M Anirudh, and J Bhuvana. 2017. Bharathi ssn@ inli-fire-2017: Svm based approach for indian native language identification. In *FIRE*.

Rupal Bhargava, Jaspreet Singh, Shivangi Arora, and Yashvardhan Sharma. 2017. Bits_pilani@ inli-fire-2017: Indian native language identification using deep learning. In *FIRE*.

Johannes Bjerva, Gintarė Grigonytė, Robert Östling, and Barbara Plank. 2017. Neural networks and spelling features for native language identification. In *Proceedings of BEA*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*.

Abderrahim Boudlal, Abdelhak Lakhouaja, Azzeddine Mazroui, Abdelouafi Meziane, MOAO Bebah, and Mostafa Shoul. 2010. Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts. In *Elsevier IACIT*.

Julian Brooke and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. In *Proceedings of LCR*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Proceedings of NeurIPS*.

Serhiy Bykh, Sowmya Vajjala, Julia Krivanek, and Detmar Meurers. 2013. Combining shallow and linguistically motivated features in native language identification. In *Proceedings of BEA*.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Najim Dehak, Pedro A Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak. 2011. Language recognition via i-vectors and dimensionality reduction. In *In Proceedings of ISCA*.

Iria del Río. 2020. Native language identification on l2 portuguese. In *Springer CPPL*.

Iria del Río, Marcos Zampieri, and Shervin Malmasi. 2018. A portuguese native language identification dataset. In *Proceedings of BEA*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Elisa Di Nuovo, Cristina Bosco, Elisa Corino, et al. 2020. How good are humans at native language identification? a case study on italian l2 writings. In *Proceedings of ICCL*.

Rod Ellis. 2015. *Understanding second language acquisition 2nd edition*. Oxford university press.

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical considerations in nlp shared tasks. In *Proceedings of EthNLP*.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for english emails. In *Proceedings of PACLING*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *the Journal of machine Learning research*, 9:1871–1874.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, et al. 2020. A report on the vardial evaluation campaign 2020. In *Proceedings of VarDial*.

Susan M Gass. 1988. Second language acquisition and linguistic theory: The role of language transfer. In *Linguistic theory in second language acquisition*, pages 384–403. Springer.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with tf-idf weighting. In *Proceedings of BEA*.

Cyril Goutte and Serge Léger. 2017. Exploring optimal voting in native language identification. In *Proceedings of BEA*.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2013. Feature space selection and combination for native language identification. In *Proceedings of BEA*.

Sylviane Granger, Estelle Dagneaux, and Fanny Meunier. 2009. The international corpus of learner english: Handbook and cd-rom, version 2. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Aman Gupta. 2018. Team webarch at fire-2018 track on indian native language identification. In *Proceedings of FIRE*.

Zhaohong Han and Terence Odlin. 2005. *Studies of fossilization in second language acquisition*, volume 14. Multilingual Matters.

John Henderson, Guido Zarrella, Craig Pfeifer, and John D Burger. 2013. Discriminating non-native english with 350 words. In *Proceedings of BEA*.

Pavel Ircing, Jan Švec, Zbyněk Zajíc, Barbora Hladká, and Martin Holub. 2017. Combining textual and speech features in the nli task using state-of-the-art machine learning techniques. In *Proceedings of BEA*.

Shin'ichiro Ishikawa. 2011. A new horizon in learner corpus studies: The aim of the icnale project. *Korea*, 404:89–168.

Royal Jain, Venkatesh Duppada, and Sushant Hiray. 2017. Seernet@ inli-fire-2017: Hierarchical ensemble for indian native language identification. In *Proceedings of FIRE*.

Scott Jarvis, Rosa Alonso Alonso, and Scott Crossley. 2019. Native language identification by human judges. *Cross-Linguistic Influence: From Empirical Evidence to Classroom Practice*, pages 215–231.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of BEA*.

Scott Jarvis and Scott A Crossley. 2012. *Approaching language transfer through text classification: Explorations in the detection based approach*, volume 64. Multilingual Matters.

Scott Jarvis and Aneta Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. Routledge.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of ACM SIGKDD*.

Dijana Kosmajac and Vlado Keselj. 2017. Dalteam@ inli-fire-2017: Native language identification using svm with sgd training. In *Proceedings of FIRE*.

Stephen Krashen. 1981. Second language acquisition. *Second Language Learning*, 3(7):19–39.

G Radha Krishna, R Krishnan, and VK Mittal. 2019. An automated system for regional nativity identification of indian speakers from english speech. In *Proceedings of IEEE INDICON*.

Artur Kulmizev, Bo Blankers, Johannes Bjerva, Malvina Nissim, Gertjan van Noord, Barbara Plank, and Martijn Wieling. 2017. The power of character n-grams in native language identification. In *Proceedings of BEA*.

Anand Kumar, Bharathi Ganesh, Ajay SG, and Soman KP. 2018. Overview of the second shared task on indian native language identification (inli). In *Proceedings of FIRE*.

Sowmya Lakshmi and BR Shambhavi. 2017. Bmsce_ise@ inli-fire-2017: A simple n-gram based approach for native language identification. In *Proceedings of FIRE*.

Wen Li and Liang Zou. 2017. Classifier stacking for native language identification. In *Proceedings of BEA*.

Zoey Liu, Tiwalayo Eisape, Emily Prud'hommeaux, and Joshua K Hartshorne. 2022. Data-driven crosslinguistic syntactic transfer in second language learning. In *Proceedings of CogSci*.

Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. A deep generative approach to native language identification. In *Proceedings of COLING*.

André Lynum. 2013. Native language identification using large scale lexical features. In *Proceedings of BEA*.

Roy C Major. 2007. Identifying a foreign accent in an unfamiliar language. *Studies in second language acquisition*, 29(4):539–556.

Shervin Malmasi. 2016. *Native Language Identification: Explorations and Applications*. Ph.D. thesis, Macquarie University.

Shervin Malmasi and Aoife Cahill. 2015. Measuring feature diversity in native language identification. In *Proceedings of BEA*.

Shervin Malmasi, Iria del Río, and Marcos Zampieri. 2018. Portuguese native language identification. In *Proceedings of PROPOR*.

Shervin Malmasi and Mark Dras. 2014a. Arabic native language identification. In *Proceedings of ANLP*.

Shervin Malmasi and Mark Dras. 2014b. Chinese native language identification. In *Proceedings of EACL*.

Shervin Malmasi and Mark Dras. 2014c. Finnish native language identification. In *Proceedings of ALTA*.

Shervin Malmasi and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of NAACL*.

Shervin Malmasi and Mark Dras. 2017a. Multilingual native language identification. *Natural Language Engineering*, 23(2):163–215.

Shervin Malmasi and Mark Dras. 2017b. Native language identification using stacked generalization. *arXiv preprint arXiv:1703.06541*.

Shervin Malmasi and Mark Dras. 2018. Native language identification with classifier stacking and ensembles. *Computational Linguistics*, 44(3):403–446.

Shervin Malmasi, Mark Dras, Mark Johnson, Lan Du, and Magdalena Wolska. 2017a. Unsupervised text segmentation based on native language characteristics. In *Proceedings of ACL*.

Shervin Malmasi, Mark Dras, and Irina Temnikova. 2015a. Norwegian native language identification. In *Proceedings of RANLP*.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017b. A report on the 2017 native language identification shared task. In *Proceedings of BEA*.

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015b. Oracle and human baselines for native language identification. In *Proceedings of BEA*.

Ilia Markov, Lingzhen Chen, Carlo Strapparava, and Grigori Sidorov. 2017. Cic-fbk approach to native language identification. In *Proceedings of BEA*.

Ilia Markov and Grigori Sidorov. 2018. Cic-ipn@ inli2018: Indian native language identification. In *FIRE (Working Notes)*.

David Martinez, Oldřich Plchot, Lukáš Burget, Ondřej Glembek, and Pavel Matějka. 2011. Language recognition in ivectors space. In *In Proceedings of ISCA*.

Seifeddine Mechti, Ayoub Abbassi, Lamia Hadrich Belguith, and Rim Faiz. 2016. An empirical method using features combination for arabic native language identification. In *proceedings of AICCSA*.

Seifeddine Mechti, Nabil Khoufi, and Lamia Hadrich Belguith. 2020. Improving native language identification model with syntactic features: Case of arabic. In *Proceedings of ISDA*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Tomoya Mizumoto, Yuta Hayashibe, Keisuke Sakaguchi, Mamoru Komachi, and Yuji Matsumoto. 2013. Naist at the nli 2013 shared task. In *Proceedings of BEA*.

Elham Mohammadi, Hadi Veisi, and Hessam Amini. 2017. Native language identification using a mixture of character and word n-grams. In *Proceedings of BEA*.

Soumik Mondal, Athul Harilal, and Alexander Binder. 2018. Corplab inli@ fire-2018: Identification of indian native language using pairwise coupling. In *Proceedings of FIRE*.

Rajesh Kumar Mundotiya, Manish Kumar Singh, and Anil Kumar Singh. 2018. Nlprl@ inli-2018: Hybrid gated lstm-cnn model for indian native language identification. In *Proceedings of FIRE*.

Hamada A Nayel and HL Shashirekha. 2017. Mangalore-university@ inli-fire-2017: Indian native language identification using support vector machines and ensemble approach. In *Proceedings of FIRE*.

Hamada A Nayel and HL Shashirekha. 2018. Mangalore university inli@ fire2018: Artificial neural network and ensemble based models for inli. In *Proceedings of FIRE*.

Yoo Rhee Oh, Hyung-Bae Jeon, Hwa Jeon Song, Yun-Kyung Lee, Jeon-Gue Park, and Yun-Keun Lee. 2017. A deep-learning based native-language classification by using a latent semantic analysis for the nli shared task 2017. In *Proceedings of BEA*.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Marius Popescu and Radu Tudor Ionescu. 2013. The story of the characters, the dna and the native language. In *Proceedings of BEA*.

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016. On the similarities between native, non-native and translated texts. In *Proceedings of ACL*.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*.

Taraka Rama and Çağrı Çöltekin. 2017. Fewer features perform well at native language identification task. In *Proceedings of BEA*.

Nikita Remnev. 2019. Native language identification for russian. In *Proceedings of ICDMW*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for esl correction tasks. In *Proceedings of ACL*.

Stian Steinbakken and Björn Gambäck. 2020. Native-language identification with attention. In *Proceedings of ICON*.

Michael Swan and Bernard Smith. 2001. *Learner English: A teacher's guide to interference and other problems*. Cambridge University Press.

Ben Swanson and Eugene Charniak. 2012. Native language detection with tree substitution grammars. In *Proceedings of ACL*.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of BEA*.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of COLING*.

D Thenmozhi, Kawshik Kannan, and Chandrabose Aravindan. 2017. Ssn_nlp@ inli-fire-2017: A neural network approach to indian native language identification. In *Proceedings of FIRE*.

D Thenmozhi, S Kayalvizhi, and Chandrabose Aravindan. 2018. A machine learning approach to indian native language identification. In *FIRE (working notes)*.

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. 2013. Identifying the l1 of non-native writers: the cmu-haifa system. In *Proceedings of BEA*.

Ludmila Tydlitátová. 2016. Native language identification of l2 speakers of czech. *Univerzita Karlova, Matematicko-fyzikální fakulta*.

Ahmet Yavuz Uluslu. 2023. Turkish native language identification. In *Proceedings of ICNLSP*.

Ahmet Yavuz Uluslu and Gerold Schneider. 2022. Scaling native language identification with transformer adapters. In *Proceedings of ICNLSP*.

Sowmya Vajjala and Sagnik Banerjee. 2017. A study of n-gram and embedding representations for native language identification. In *Proceedings of BEA*.

Matias Johansen Vian. 2023. A study of transformers for cross-corpus native language identification. Master's thesis, NTNU.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of ALTA*.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of EMNLP*.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2012. Exploring adaptor grammars for native language identification. In *Proceedings of EMNLP*.

Ching-Yi Wu, Po-Hsiang Lai, Yang Liu, and Vincent Ng. 2013. Simple yet powerful native language identification on toefl11. In *Proceedings of BEA*.

Marcos Zampieri, Alina Maria Ciobanu, and Liviu P. Dinu. 2017. Native language identification on text and speech. In *Proceedings of BEA*.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2024. Language variety identification with true labels. In *Proceedings of LREC-COLING*.

Wei Zhang and Alexandre Salle. 2023. Native language identification with large language models. *arXiv preprint arXiv:2312.07819*.

Marc A Zissman. 1996. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on speech and audio processing*, 4(1):31.

Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2016. Tweetlid: a benchmark for tweet language identification. *Language Resources and Evaluation*, 50:729–766.