Multilingual Native Language Identification with Large Language Models

Dhiman Goswami[1], Marcos Zampieri[1], Kai North[2]
Shervin Malmasi[3], Antonios Anastosopoulos[1]
[1]George Mason University, USA, [2]Cambium Assessment, USA
[3]Amazon.com, Inc. USA
dgoswam@gmu.edu

Abstract

Native Language Identification (NLI) is the task of automatically identifying the native language (L1) of individuals based on their second language (L2) production. The introduction of Large Language Models (LLMs) with billions of parameters has renewed interest in text-based NLI, with new studies exploring LLM-based approaches to NLI on English L2. The capabilities of state-of-the-art LLMs on non-English NLI corpora, however, have not yet been fully evaluated. To fill this important gap, we present the first evaluation of LLMs for multilingual NLI. We evaluated the performance of several LLMs compared to traditional statistical machine learning models and language-specific BERT-based models on NLI corpora in English, Italian, Norwegian, and Portuguese. Our results show that fine-tuned GPT-4 models achieve state-of-the-art NLI performance.

1 Introduction

Individuals proficient in a language have the ability to identify accent patterns in non-native speech (Major, 2007). Automatically identifying a speaker's native language (L1) when speaking a second language (L2) on the basis of pronunciation, stress, and prosodic patterns has been substantially explored in speech-based NLI (Krishna et al., 2019). Similarly, in text-based NLI, linguistic patterns common to an individual's L1 such as word choices, syntax, and spelling, can be recognized in texts written in a given L2. Computational models can be then trained on texts authored by non-native speakers to learn distinctive properties of their L1, aiming to identify the writer's mother tongue (Malmasi, 2016).

The underlying assumption in NLI is that the native language influences Second Language Acquisition (SLA) and production, a phenomenon known as cross-linguistic influence or language transfer (Krashen, 1981; Ellis, 2015). Language transfer results in L1 features manifesting in L2 production, allowing computational models to recognize patterns shared by speakers of the same L1 when communicating in a given L2. Text-based NLI has numerous important applications such as serving as a corpus-driven approach for SLA (Jarvis and Crossley, 2012) and enabling the development of effective L2 teaching materials and computer-aided language learning (CALL) software. Additionally, NLI has been shown to improve NLP systems dealing with texts from non-native speakers, contributing to tasks like author profiling, forensics, spam and phishing detection (Malmasi et al., 2017).

As evidenced by a recent survey (Goswami et al., 2024), traditional statistical models such as Support Vector Machines (SVMs) trained on n-grams as features have historically delivered the best performance for text-based NLI. A few recent studies (Zhang and Salle, 2023; Ng and Markov, 2024), however, have shown that fine-tuned LLMs such as GPT-4 deliver state-of-the-art performance for English NLI. A key limitation of these studies, as discussed by Ng and Markov (2024) is the lack of evaluation of LLMs for languages other than English. To address this important gap in the literature, we propose the first multilingual evaluation of LLMs in NLI. We evaluate various LLMs, in a zero-shot and fine-tuned setting, on corpora containing English, Italian, Norwegian, and Portuguese L2 production.

We investigate two research questions (RQs):

- **RQ1:** How effectively can LLMs identify L1s across NLI datasets in English and other languages?
- **RQ2:** To what extent does task-specific fine-tuning improve the performance of LLMs compared to zero-shot prompting across different languages?

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 193–199
April 30 - May 1, 2025 ©2025 Association for Computational Linguistics

2 Related Work

The aforementioned survey by Goswami et al. (2024) presents a comprehensive account of text-based NLI, covering more than 100 papers on the topic. It describes studies that use a variety of features such as word n-grams (Gebre et al., 2013), part-of-speech tags (Wong et al., 2012), and syntactic features (Wong and Dras, 2011; Mechti et al., 2020). The survey also covers computational models widely employed in text-based NLI from statistical classifiers like SVMs (Jarvis et al., 2013; Goutte et al., 2013) and Logistic Regression (Tsvetkov et al., 2013; Popescu and Ionescu, 2013; Gupta, 2018) to deep learning architectures (Ajees and Idicula, 2018; Lotfi et al., 2020; Uluslu and Schneider, 2022) and LLMs (Zhang and Salle, 2023). In addition, it reviews shared tasks organized on the topic that provided important benchmark text-based datasets (Tetreault et al., 2013; Malmasi et al., 2017; Soman, 2018).

The findings described in Goswami et al. (2024) reveal that until recently, approaches that combined statistical classifiers with feature engineering achieved state-of-the-art performance on text-based NLI while deep learning architectures achieved limited success. Recent studies, however, have showed that the latest generation of LLMs, most notably GPT-4, are able to outperform statistical and previous neural models (Zhang and Salle, 2023) particularly when such models are fined-tuned for text-based NLI (Ng and Markov, 2024).

The majority of studies referenced here, including recent studies on LLM architectures (Ng and Markov, 2024), only address English NLI. This is due to the wider availability of English L2 corpora compared to other languages including widely-used learner corpora such as ICLE (Granger et al., 2009), TOEFL11 (Blanchard et al., 2013), and ICNALE (Ishikawa, 2011). Multiple multilingual studies have been conducted that describe data and approaches to text-based NLI in other L2s. This includes studies on Arabic (Malmasi and Dras, 2014a; Ionescu, 2015; Bassas and Kübler, 2024), Chinese (Malmasi and Dras, 2014b), Czech (Tydlitátová, 2016), Finish (Malmasi and Dras, 2014c), Norwegian (Malmasi et al., 2015), Portuguese (Malmasi et al., 2018; del Río, 2020), and Turkish (Uluslu and Schneider, 2023).

To the best of our knowledge, all text-based NLI studies on L2 other than English employed traditional machine learning models combined with feature engineering or early deep learning approaches. The use of LLMs for L2s other than English remains unexplored. Our work fills this gap by presenting the first multilingual evaluation of LLMs in text-based NLI on four languages and five datasets.

3 Data

In this study we use five NLI corpora in English, Italian, Norwegian, and Portuguese. NLI corpora, and learner corpora in general, are only available for English and a few other high-resource languages (Malmasi, 2016; Goswami et al., 2024) which limits the choice of languages we can study. With the goal of carrying out a multilingual evaluation, we choose Italian, Norwegian, and Portuguese due to the availability of suitable corpora.

Data Splits For TOEFL11 and NLI-PT we follow pre-defined training, development, and testing split from prior work (Tetreault et al., 2012; Malmasi et al., 2018). For all other corpora, we use a random label wise 80%-10%-10% split for training, development, and testing. To ensure comparability of results, we use the same splits across the different experiments presented in the paper. Brief descriptions of the five corpora are presented next.

English - FCE and TOEFL11 For L2 English, we use FCE and TOEFL11. FCE contains 1,244 exam scripts extracted from the Cambridge Learner Corpus (CLC) and written by candidates who took the Cambridge ESOL First Certificate in English (FCE) in 2000 and 2001 (Malmasi, 2016). It includes the following L1s: Spanish, French, Korean, Russian, Japanese, Turkish, Polish, Italian, Greek, German, Portuguese, Chinese, Catalan, Thai, Swedish, and Dutch. TOEFL 11 (Tetreault et al., 2012) is a dataset of essays written by speakers of 11 L1s: Arabic, German, French, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish and Chinese. Following the split by Tetreault et al. (2012) we use 1,100 essays for each L1 with 900 for training, 100 for development, and 100 for testing.

Italian - VALICO For Italian, we use VALICO (Corino et al., 2017), the *Varieta di Apprendimento deLlla Lingua Italiana Corpus Online*, i.e. Online Corpus of Learner Varieties of Italian. VALICO contains 2,531 texts written by L1 speakers of Albanian, Chinese, Czech, English, French, German, Hindi, Japanese, Polish, Portuguese, Romanian, Russian, Serbian, Spanish.

Norwegian - ASK For Norwegian, we use ASK (Tenfjord et al., 2006), the *Andrespråkskorpus*, i.e. Second Language Corpus. It features essays written in Norwegian Bokmål as part of an exam in Norwegian as a second language. It covers 2,158 essays written by L1 speakers of Albanian, Dutch, English, German, Polish, Russian, Serbian, Somali, Spanish, and Vietnamese.

Portuguese - NLI-PT For Portuguese, we acquire NLI-PT (del Río et al., 2018). NLI-PT is a corpus collected from three learner corpora of Portuguese: (i) COPLE2; (ii) Leiria corpus, and (iii) PEAPL2. It contains written productions from learners of European Portuguese with different proficiency levels and L1s. We use 1,075 texts written by L1 speakers of Chinese, Spanish, English, Italian, and German and the same train, development, and test split as in Malmasi et al. (2018).

4 Models

Statistical Machine Learning Ensemble We trained a Logistic Regression (LR) and an SVM classifier on POS n-grams of $n \in [1, 4]$. The data was normalized and its dimensionality was reduced using TruncatedSVD and PCA. We then combine the LR and SVM models in a majority voting ensemble (Malmasi and Dras, 2017). We refer to this model as ML Ensemble.

Transformers We fine-tune two multilingual models for the four languages, namely mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). We also fine-tune several language specific models. For English we use BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), for Italian we use italianBERT (Dbmdz, 2020), for Norwegian we use norBERT (Samuel et al., 2023), and for Portuguese we use BERTimbau (Souza et al., 2020). We use learning rate 1e-5 for all models. The hyperparameters for the transformer models for all corpora are presented in Table 1.

Dataset	Epochs	Batch Size
FCE	5	8
TOEFL11	3	16
VALICO	10	16
ASK	5	16
NLI-PT	5	8

Table 1: Hyperparameters for BERT-based transformers and Flan-T5.

LLM Prompting We use FLAN-T5 (Chung et al., 2024) and GPT-4 (Achiam et al., 2023) for zero-shot prompting. We also carried out preliminary experiments with various 7 billion parameter models (e.g., Mistral-7B (Jiang et al., 2023)) which obtained much lower performance overall and therefore have not been included in our experiments. A sample LLM prompt used in our experiments is presented below.

```
Role (system): You are a forensic
linguistics expert that reads <L2 Language>
texts written by non-native authors in
order to classify the native language of the
author as one of: <List of L1s>. The output
will be the short form of the languages
in this list - <label>. Use clues such
as spelling errors, word choice, syntactic
patterns, and grammatical errors to decide.
DO NOT USE ANY OTHER CLASS.
IMPORTANT: Do not classify any input as
<L2 Language>. <L2 Language> is an invalid
choice.
Role (user): <a text written by a
non-native speaker>
```

LLM Fine-tuning We further fine-tune FLAN-T5 and GPT-4 for all datasets. For FLAN-T5, we have used the same epochs and batch size presented in Table 1. For GPT-4, we use the API provided OpenAI.[1] The data gets validated and an optimal set of hyperparameters are automatically fixed for fine-tuning. The hyperparameters of GPT-4 fine-tuning for all the datasets are given in Table 2 while the learning rate for all languages is 2e-5.

Dataset	Epochs	Batch Size
FCE	3	2
TOEFL11	2	16
VALICO	3	4
ASK	3	3
NLI-PT	3	2

Table 2: Hyperparameter for GPT-4 Fine-Tuning.

5 Results

We present the results for all languages in terms of accuracy and Macro F1, which is the standard in prior work (Malmasi, 2016; Goswami et al., 2024). The results are presented along with a random and a majority class baseline for comparison. Finally, to ensure a fair and comparable analysis across all

[1] https://platform.openai.com/finetune/

experiments, we evaluate all models on the same test sets for each particular corpus.

5.1 English

The results for English are presented in Table 3. We observe that all models achieve performance significantly higher than the two baselines provided. The results across the two NLI datasets demonstrate that LLMs, and in particular GPT-4, achieve state-of-the-art performance in text-based NLI when fine-tuned for the task. As shown in Table 3, fine-tuned GPT-4 achieves the highest F1 scores on both corpora with 0.82 for FCE and 0.92 for TOEFL11.

It is also worth noting that the the GPT-4 prompting results for TOEFL11 are unusually high compared to the results obtained by this model on the other four corpora. This is in line with the results reported by (Zhang and Salle, 2023) on TOEFL11. The high results suggest that model may have seen instances from this dataset indicating potential data contamination.

	FCE		**TOEFL11**	
Models	**Acc.**	**F1**	**Acc.**	**F1**
Random Baseline	0.06	0.06	0.10	0.10
Majority Baseline	0.14	0.04	0.09	0.02
ML Ensemble	0.47	0.46	0.84	0.82
BERT	0.25	0.25	0.68	0.68
mBERT	0.27	0.27	0.67	0.66
RoBERTa	0.29	0.28	0.71	0.71
XLM-R	0.33	0.32	0.63	0.62
FLAN T5 Prompt	0.38	0.36	0.32	0.32
GPT-4 Prompt	0.39	0.39	0.83	0.83
FLAN-T5 FT	0.37	0.36	0.73	0.73
GPT-4 FT	**0.83**	**0.82**	**0.92**	**0.92**

Table 3: Model results and baselines for English in terms of Accuracy (Acc.) and Macro F1 (F1). "Prompt" indicates zero-shot prompting, "FT" indicates fine-tuning.

Another key finding is that LLM fine-tuning outperforms all other models by a substantial margin. The ML ensemble, achieves 0.82 F1 for TOEFL11 lagging significantly behind the fine-tuned GPT-4 model. Another notable trend is that fine-tuning drastically improve LLM performance over zero-shot prompting. For example, on TOEFL11, while GPT-4 zero-shot gets 0.83 F1, fine-tuning boosts the performance to 0.92 F1. Finally, when comparing multilingual and language-specific transformer models, we obtain mixed results. On TOEFL11, monolingual models like RoBERTa outperform multilingual ones, while on FCE, multilingual XLM-R performs better than RoBERTa.

5.2 Italian, Norwegian, and Portuguese

Results for Italian, Norwegian, and Portuguese are presented in Table 4. Similarly to what we observed for English, we see a significant effect of task fine-tuning over zero-shot prompting on the LLMs performance. This is evidenced by the GPT-4 performance which, for Italian, achieves 0.78 F1 score when fine-tuned and 0.31 F1 when prompting. A similar trend is observed for Norwegian and Portuguese. We observe that the zero-shot results are much lower for Italian, Norwegian, and Portuguese when compared to English. This is somewhat expected as LLMs have shown to possess greater capabilities for English compared to all other languages (Minaee et al., 2024).

We see that the ML Ensemble outperforms all of the Transformer-based small LMs for all languages. This confirms the findings of related studies as discussed in a recent survey (Goswami et al., 2024). Finally, with the exception of norBERT for Norwegian, we see that language-specific transformers such as italianBERT and BERTimbau outperform the multilingual models mBERT and XLM-R. The reinforces the mixed results on language-specific vs. multilingual transformer models we described for English.

6 Conclusion and Future Work

This paper presented the first evaluation of LLMs on multilingual text-based NLI, experimenting with four languages and five corpora. Our results indicate that larger task fine-tuned LLMs, such as GPT-4, deliver state-of-the-art performance for text-based NLI in the four languages studied. This finding is in line with prior results obtained for English NLI (Ng and Markov, 2024).

We further observed that for non-English languages, zero-shot LLM prompting approaches are generally outperformed by BERT-based and statistical ML approaches. This is likely a limitation of the LLMs we leveraged here, as they are mostly focused on English.

We conclude the paper by revisiting the two RQs below and presenting avenues for future work.

RQ1: How effectively can LLMs identify L1s across NLI datasets in English and other languages?

	Italian		Norwegian		Portuguese	
Model	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
Random Baseline	0.09	0.10	0.10	0.11	0.13	0.14
Majority Baseline	0.16	0.04	0.13	0.03	0.27	0.11
ML Ensemble	0.66	0.63	0.76	0.76	0.59	0.59
italianBERT	0.45	0.42	-	-	-	-
norBERT	-	-	0.67	0.67	-	-
BERTimbau	-	-	-	-	0.56	0.55
mBERT	0.48	0.44	0.43	0.40	0.57	0.57
XLM-R	0.43	0.36	0.42	0.39	0.32	0.30
FLAN T5 Prompt	0.28	0.27	0.37	0.36	0.30	0.29
GPT-4 Prompt	0.31	0.31	0.52	0.51	0.45	0.36
FLAN T5 FT	0.62	0.57	0.73	0.72	0.45	0.42
GPT-4 FT	**0.79**	**0.78**	**0.92**	**0.92**	**0.86**	**0.86**

Table 4: Model results and baselines for Italian, Norwegian and Portuguese in terms of Accuracy and Macro F1 (F1). "Prompt" indicates zero-shot prompting while "FT" indicates fine-tuning.

> **RQ1 Results:** Our evaluation of LLM zero-shot prompting indicates that LLMs have very little knowledge of NLI for the four languages and five corpora explored. A notable exception is TOEFL11, the most popular NLI corpus available, for which the results obtained using GPT-4 were very high using zero-shot prompting. This seems to indicate potential data contamination. When fine-tuned, we observed that LLM results have significantly increased (see RQ2 Results). Finally, for all languages, statistical ML classifiers obtained performance superior to several transformers and LLM prompting.

RQ2: To what extent does task-specific fine-tuning improve the performance of LLMs compared to zero-shot prompting across different languages?

> **RQ2 Results:** When fine-tuned to the task, GPT-4 achieves state-of-the-art performance for all four languages and five corpora explored. Furthermore, we observe that the performance gap between zero-shot and fine-tuning is much smaller for English compared to the other three languages. This provides further evidence of the ability of LLMs to better deal with English data compared to all other languages.

In future work, we would like to use the output of these classifiers to carry out a cross-lingual study of L1 to L2 transfer. This has been done extensively in the past using statistical classifiers (Jarvis and Crossley, 2012; Bykh and Meurers, 2014; Malmasi, 2016). We believe that the models used in our experiments may reveal interesting linguistic patterns being transferred from L1 that may generalize to various L2s in terms of spelling, word choices, and syntax.

Limitations

We hope the results presented in this paper motivate further research in multilingual NLI. The limitations of this work are related to the choice of languages and models. With respect to the languages, there are unfortunately very few corpora available for NLI which limits the choice of languages we can study. We hope our findings motivate researchers to create new NLI corpora for languages other than English and, in particular, for low-resource languages. All four languages that we studied are considered to be high-resourced. Finally, with respect to the models, we would like to investigate the performance of recently released LLMs such as Gemma, as in Ng and Markov (2024), on multilingual text-based NLI.

Acknowledgments

We thank the anonymous reviewers for the feedback provided. We further thank the creators of the five datasets used in this study for making the data available to the community. Antonios Anastasopoulos is generously supported by the National Science Foundation under award IIS-2327143.

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*.

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

Yasmeen Bassas and Sandra Kübler. 2024. Investigating linguistic features for arabic nli. In *Proceedings of ArabicNLP*.

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

Serhiy Bykh and Detmar Meurers. 2014. Exploring syntactic features for native language identification: A variationist perspective on feature encoding and ensemble optimization. In *Proceedings of COLING*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25:1–53.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Elisa Corino, Carla Marello, Simona Colombo, et al. 2017. *Italiano di stranieri. I corpora VALICO e VINCA*, volume 6. Guerra.

Dbmdz. 2020. BERT-base italian cased model. https://huggingface.co/dbmdz/bert-base-italian-cased.

Iria del Río. 2020. Native language identification on l2 portuguese. In *Proceedings of PROPOR*.

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*.

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

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

Dhiman Goswami, Sharanya Thilagan, Kai North, Shervin Malmasi, and Marcos Zampieri. 2024. Native language identification in texts: A survey. In *Proceedings of NAACL*.

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 *FIRE*.

Radu Tudor Ionescu. 2015. A fast algorithm for local rank distance: Application to arabic native language identification. In *Proceedings of ICONIP*.

Shin'ichiro Ishikawa. 2011. A new horizon in learner corpus studies: The aim of the icnale projects. In *In G. Weir, S. Ishikawa, and K. Poonpon, editors, Cor56 pora and Language Technologies in Teaching, Learning and Research*. University of Strathclyde Publishing.

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 Detectionbased Approach*. Multilingual Matters.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Stephen Krashen. 1981. Second language acquisition. *Second Language Learning*.

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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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

Roy C Major. 2007. Identifying a foreign accent in an unfamiliar language. *Studies in second language acquisition*.

Shervin Malmasi. 2016. *Native language identification: explorations and applications*. Ph.D. thesis, Macquarie University, Faculty of Science and Engineering, Department of CLT.

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 EMNLP (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. 2017. Multilingual native language identification. *Natural Language Engineering*.

Shervin Malmasi, Mark Dras, and Irina Temnikova. 2015. 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. 2017. A report on the 2017 native language identification shared task. In *Proceedings of BEA*.

Seifeddine Mechti, Nabil Khoufi, and Lamia Hadrich Belguith. 2020. Improving native language identification model with syntactic features: Case of arabic. In *SPring ISDA 2018*.

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*.

Yee Man Ng and Ilia Markov. 2024. Leveraging open-source large language models for native language identification. In *Proceedings of VarDial*.

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

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of NoDaLiDa*.

KP Soman. 2018. Overview of the second shared task on indian native language identification (inli). In *CEUR workshop proceedings*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Proceedings of BRACIS*.

Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006. The ask corpus–a language learner corpus of norwegian as a second language. In *Proceedings of LREC*.

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*.

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. Bachelors thesis, Charles University.

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

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

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*.

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