# Native language Identification for Arabic Language Learners using Pre-trained Language Models

**Mohamed Amine Cheragui**[1], **Mourad Abbas** [2] and **Mohammed Mediani**[3]
[1] Mathematics and Computer Science Department Ahmed Draia University, Adrar - Algeria
[2] High Council of Arabic - Algeria
[3] College of Information Technology, United Arab Emirates University Al-Ain, UAE
m_cheragui@univ-adrar.edu.dz
m_abbas04@yahoo.fr
mohammed.mediani@uaeu.ac.ae

## Abstract

In this paper, we conduct an empirical study designed to systematically evaluate the efficacy of deep learning approaches in Native Language Identification (NLI) for native and non-native Arabic speakers. Specifically, we utilize three models: CAMeLBERT, AraBERTv0.2, and ARBERTv2. Our analysis is structured around two classification scenarios: binary classification and multi-class classification. This methodological framework allows us to comprehensively assess the performance of each model for the designated task.

## 1 Introduction

Native Language Identification is a specialized area within natural language processing (NLP) focused on automatically determining an individual's first language (L1) or mother tongue based on their written or spoken text in a second language (L2). This field involves the analysis of various linguistic features—including vocabulary usage, syntax, and stylistic patterns—to deduce the most likely native language of a writer or speaker. This process is predicated on the hypothesis that linguistic characteristics of the mother tongue often manifest in the acquisition and use of a second language, a phenomenon known as language transfer (Zampieri et al., 2017).

NLI offers a range of practical applications across diverse fields: Authorship Identification (Authorship Attribution) (Jarvis and Paquot, 2015), Author Profiling (Estival et al., 2007), Forensic Linguistics, (Mohammadi et al., 2017), Human-machine voice interface applications (Qian et al., 2017), Second Language Acquisition (SLA) (Malmasi and Dras, 2017b), Educational Technology Development (Laufer and Girsai, 2008), Marketing (Chen et al., 2017), and Security (Malmasi and Dras, 2017a).

In the literature, most research on NLI has focused on integrating linguistic features with machine learning methods (Tetreault et al., 2013). Key linguistic features analyzed include part-of-speech (POS) tagging (Gebre et al., 2013), character n-grams (Kulmizev et al., 2017), spelling errors (Kyle et al., 2015), and syntactic features (Wong and Dras, 2011). Commonly employed machine learning techniques in this domain include Naïve Bayes (NB) and Support Vector Machines (SVM). This combination leverages both detailed linguistic analysis and advanced computational models to effectively predict the native language of individuals from their second language texts.

The objective of our study is to conduct a series of experiments to investigate the efficacy of deep learning approaches in NLI for Arabic language learners. We explore this through two classification scenarios: binary classification and multi-class classification. To this end, we employ three models based on Bidirectional Encoder Representations from Transformers (BERT): CAMeLBERT (Inoue et al., 2021), AraBERTv0.2 (Antoun et al., 2020), and ARBERTv2 (Abdul-Mageed et al., 2021). These models are specifically implemented to assess the contribution of deep learning techniques in accurately identifying the native languages of Arabic language learners.

The structure of this paper is organized as follows: Section 2 reviews related work in NLI, offering background and context for our study. Section 3 describes the methodology and datasets used in our experiments, detailing the computational models, analysis techniques, and evaluation of each model's performance across various classification scenarios. Section 4 discusses the findings. Finally, Section 5 concludes the paper and suggests potential directions for future research in this field.

## 2 Literature Review

Like all other topics specific to natural language processing, research in NLI was focused essentially on learning English. However, in recent years a number of studies have focused on other languages as Chinese, Norwegian, Portuguese and Arabic.

### 2.1 English Learning Language

(Tetreault et al., 2012) conducted a pioneering study on the use of classifier ensembles for NLI. The study employed an ensemble of logistic regression learners, utilizing a diverse set of features including character and word n-grams, function words, parts of speech, spelling errors, and writing quality markers. For syntactic features, they explored the use of Tree Substitution Grammars and dependency features obtained using the Stanford parser. They also proposed incorporating language models into NLI and used language model perplexity scores based on lexical 5-grams from each language in their corpus. The ensemble model achieved accuracies, with 90.1% on the ICLE (Granger et al., 2009) and 80.9% on the TOEFL11 corpus (Blanchard et al., 2013), respectively.

(Lotfi et al., 2020) proposed a deep generative language modelling (LM) approach to NLI. Their approach is to fine-tune a GPT-2 model separately on texts written by the authors with the same L1, and assigning a label to an unseen text based on the minimum LM loss with respect to one of these fine-tuned GPT-2 models. They evaluated their approach using two datasets, TOEFL11 and ICLE, achieving an accuracy of 86.6% and 94.2% respectively.

(Uluslu and Schneider, 2022) described ProDAPT, transformer adapters based on deep generative model, which is considered as an alternative lightweight fine-tuning strategy that achieves equal performance to full fine-tuning on most tasks. In terms of performance, their model achieved 82.4% accuracy on TOEFL11 corpus.

### 2.2 Arabic Learning Language

(Malmasi and Dras, 2014) presented the first application for NLI to Arabic learners, based on a supervised multi-class classification approach, by combining three syntactic features (CFG production rules, Arabic function words and Part-of-Speech n-grams. To perform multi-class classification, they used SVM. The system achieves an accuracy of 41% on ALC Corpus.

(Mechti et al., 2020) studied the impact of automatic classification using some data statistically extracted from a source corpus, to detect the mother tongue of Arabic learners. They combined three syntactic features which are: Part of speech n-grams, function words and context-free grammar production rules. For the classification, the LIBSVM2 was used, as variant of SVM. For training and evaluation, they opted for Arabic Learner Corpus, in which their model obtained an accuracy of 45%.

(Ionescu, 2015) presented a study based on a machine learning method that works at the character level, using a kernel based on Local Rank Distance (LRD). The resulting model of this combination was trained and tested on ALC, obtained an accuracy score of 50.1%.

### 2.3 Other Learning Languages

(Malmasi et al., 2015) proposed NLI experiments on Norwegian language, by employing a supervised multi-class classification approach, which takes into consideration three syntactic feature types: function Words, part-of-Speech n-grams and mixed POS-function word n-grams. As a dataset for training and evaluation they used the ASK Corpus (Tenfjord et al., 2006). The model achieved an accuracy score of 78.6%.

(Remnev, 2019) developed a model for Russian Native Language Identification, based on the support vector method and the TF-IDF metric. To train and evaluate the proposed model, he used the Russian Learner Corpus. In terms of performance, the adopted approach achieved an accuracy score of 80%. (Malmasi et al., 2018) presented a study about native Language Identification for learners of Portuguese (as L2 Language). The used approach is a combination of linguistic features and Machine Learning. The authors defined three features which are: Function words, Context-free grammar production rules and Part-of-Speech (POS) tags. They also utilized a standard multi-class classification approach, by using linear Support Vector Machines. For the dataset, they used NLI-PT (del Río Gayo et al., 2018). The proposed model attained an accuracy of 54.1% .

(del Río, 2020) investigated the impact of different linguistic features in NLI for L2 Portuguese. For that, she defined two types of lexical features: one includes all the words in the text, and the other one includes all the words except nouns and adjectives. In addition, other morphological and syntactic features have been used, including: POS, context-free grammar (CFG) production rules and dependency triplets. For the experiment, she used 04 classifiers,

which are: Multinomial Logistic Regression, SVM, Ridge Regression and Multi-Layer Perceptron classifier, which have been trained and tested on the NLI-PT dataset. In terms of performance, the MLP classifier achieved the best accuracy of 66%.
(Uluslu, 2023) presented an application of NLI specifically for Turkish language learners. The approach employed a combination of three syntactic features: Context-Free Grammar (CFG) production rules, part-of-speech n-grams, and function words. The study used a standard supervised multi-class classification method, where a linear Support Vector Machine (SVM) was applied for classification. Feature vectors were created using a TF-IDF weighting scheme. The Turkish Learner Corpus (TLC) (Golynskaia, 2022), was utilized to evaluate the system's performance. By combining the three features, the proposed system achieved an accuracy score of 44.2%.

# 3 Experimental Methodology

## 3.1 Data

For our experiment, we used the Arabic Learner Corpus (ALC) (Alfaifi et al., 2014). The Corpus has been used for various studies in language learning and computational linguistics focusing on Arabic. It comprises a collection of written and spoken materials produced by learners of Arabic, which are used for different types of linguistic research and language teaching tool development. The dataset was compiled during the years 2012 and 2013. It comprises 282,732 words and consists of 1585 texts, encompassing both written and spoken content. These texts were generated by a total of 942 students learning Arabic, representing 67 nationalities and originating from 66 distinct mother tongue backgrounds. In addition, ALC includes 26 variables as metadata elements, 12 for the learner and 14 for the text.

## 3.2 Models

The aim of our research is to examine the impact of various pre-trained Arabic BERT models by exploring different combinations of classification task related to native language identification.
To achieve this, we fine-tuned 03 models, including AraBERTV0.2, ARBERTv2 and CAMeLBERT Using Arabic Learner Corpus. Each model was used to execute 02 Scenarios, which are: Binary classification and Multi-class classification.
Our choice of these models was made for a number of reasons: they have been specifically pre-trained on large-scale Arabic corpora, which helps them capture the nuances and intricacies of Arabic. They have demonstrated competitive performance on various NLP tasks (Sentiment Analysis, Language Identification, Named Entity Recognition, Fake News Detection, etc). Their architectures and training procedures are designed to achieve state-of-the-art results on a range of Arabic language understanding tasks, making them suitable choices for classification tasks as well. These models often come in different version (Large/base) and variants (MSA/Dialect).

It's also important to mention that even though the 03 models were developed based on the same architecture (BERT), there are a number of distinguishing features.

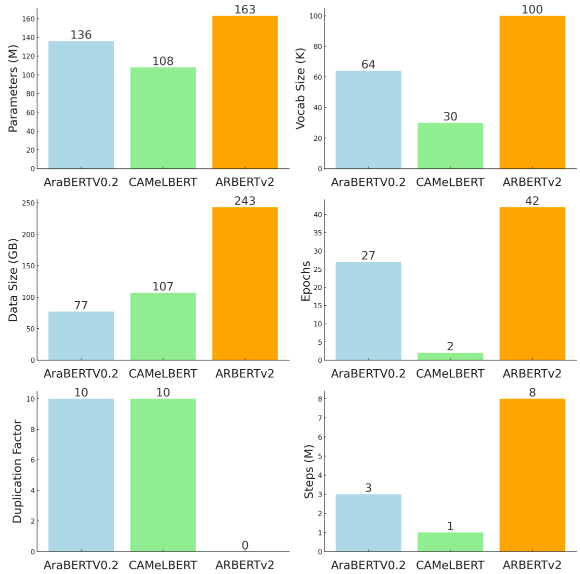


Figure 1: Configurations of used models.

| Parameter | Value |
|---|---|
| Epochs | 05 |
| Batch | 08 |
| Learning rate | 4.87 e-5 |
| weight decay | 0.01 |
| seed | 20 |

Table 1: Hyper-parameters values.

### 3.2.1 Binary Classification

Binary classification is a fundamental task in machine learning where the goal is to classify input data into one of two possible categories or classes (Er et al., 2016). To do this, We carried out two experiments, the first concerning the identification of Arabic as the mother tongue of learners, in order to fine-tuning our models, we have divided our dataset into two distinct categories. The first category is labelled "1", which concerns texts whose authors native language is Arabic. The rest of the texts constituting the second category will be labelled "0" (Table 2 summarises the different test results). The second experiment mirrors the first, but in this instance, we handle each of the six languages individually, applying the same process to each one (the result is given in the table 3).

### 3.2.2 Multi-class Classification

Multi-class classification involves classifying data into more than two groups/categories (Fields et al., 2024). Unlike binary classification, where the model is trained to predict only one of the two classes of an item, a multi-class classifier is trained to predict one of the three or more classes of an item. In our case, we set up two experiments, the first for detecting Arabic language learners' mother tongues and the second dealing with the same task based on level of study.

- **Multi-class classification for detecting Arabic learners' mother tongue**: The corpus comprises 66 distinct mother tongue representations. However, the number of representative texts varies from one Mother tongue to another, so we kept only languages with a good quantitative representation in terms of texts, as shown in figure 2. The results of this experiment are given in table 3.

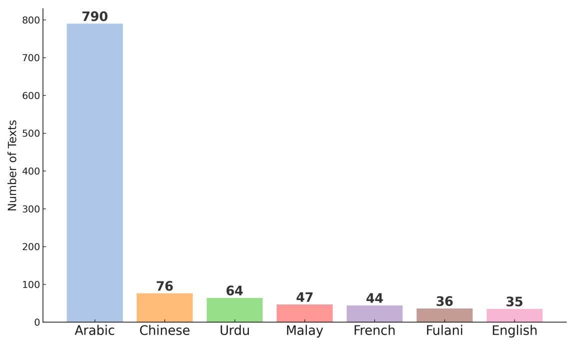


Figure 2: Number of texts produced by Arabic language learners with a mother tongue other than Arabic (ALC corpus).

- **Multi-class classification based on Level of Study**: The ALC contains 05 categories of learners according to their level of study: secondary school, general language course, diploma programme (advanced language course), Bachelor degree and Master degree. Learners of both the Bachelor degree and Master degree were majoring in Arabic. Figure 3 gives an estimate of the percentage of each level of study in the ALC. For results, table 5 presents a global view of the performance of three models in detecting the mother tongue based on levels of study, offering a comparative understanding of their effectiveness in this classification task. The Table 6 offers nuanced insights into its ability of the CAMeLBERT model to capture specific mother tongue differences at each level, providing a more refined understanding of its classification precision in this context.

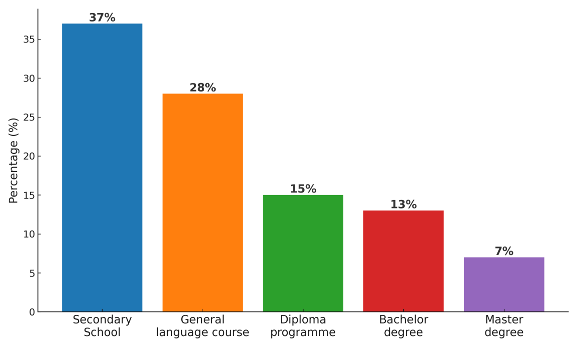


Figure 3: Corpus distribution by Level of Study.

## 4 Discussion

After reviewing the results, we found that the three models achieved good results for binary classification, but in the Multi-class classification there was a significant decrease in the effectiveness of the models, which is probably due to the fact that deep learning models such as CAMeLBERT, AraBERTv0.2 and ARBERTv2 have differences in performance when applied to multi-class classification task due to several reasons:

- Model structure and training data: These models, being variants of BERT (bi-directional encoding representations of transforms), are primarily designed to capture complex patterns in text through deep bi-directional representations. However, the effectiveness of these models is highly dependent on the quality and diversity of the training data. For Arabic with many dialects and a rich morphological

| Metrics | CAMeLBERT | AraBERTv0.2 | ARBERTv2 |
|---|---|---|---|
| Accuracy | 97.71% | 97.26% | 96.34% |
| Precision | 96.51% | 95.62% | 94.37% |
| Recall | 97.14% | 96.84% | 95.57% |
| F1 | 96.82% | 96.21% | 94.95% |

Table 2: Binary Classification, One-versus-All (Arabic versus the six other languages).

| Metrics | CAMeLBERT | AraBERTv0.2 | ARBERTv2 |
|---|---|---|---|
| Chinese | 98.32% | 96.14% | 94.43% |
| Urdu | 97.66% | 95.89% | 95.09% |
| Malay | 96.15% | 97.23% | 95.77% |
| French | 97.73% | 98.02% | 96.41% |
| Fulani | 98.18% | 97.86% | 95.31% |
| English | 97.06% | 96.23% | 93.86% |

Table 3: Binary Classification One-versus-One (Arabic/Non-Arabic).

| Metrics | CAMeLBERT | AraBERTv0.2 | ARBERTv2 |
|---|---|---|---|
| Accuracy | 87.21% | 83.10% | 81.27% |
| Precision | 64.74% | 33.49% | 28.67% |
| Recall | 61.74% | 40.59% | 35.91% |
| F1 | 60.43% | 36.45% | 30.58% |

Table 4: Multi-class detection of mother tongue learners'.

| Metrics | CAMeLBERT | AraBERTv0.2 | ARBERTv2 |
|---|---|---|---|
| Accuracy | 80.82% | 75.79% | 74.42% |
| Precision | 80.00% | 77.74% | 62.92% |
| Recall | 63.22% | 50.02% | 48.41% |
| F1 | 66.13% | 53.55% | 50.86% |

Table 5: Global view on multi-class classification performance based on Level of study using CAMeLBERT, AraBERTv0.2 and ARBERTv2.

| Metrics | Precision | Recall | F1 Score |
|---|---|---|---|
| Secondary school | 58.64% | 40.44% | 48.45% |
| General language course | 60.08% | 45.58% | 52.19% |
| Diploma programme | 62.12% | 49.24% | 56.78% |
| Bachelor degree | 78.45% | 54.97% | 63.32% |
| Master degree | 80.25% | 57.48% | 66.45% |

Table 6: Detailed scores for multi-class classification based on Level of study using CAMeLBERT.

structure, models trained on Standard Arabic may not perform well when faced with dialect-related variations unless they are specifically tuned to diverse datasets that include such variations.

- Task complexity: Multi-class classification task are inherently more complex than binary classification. In multi-class classification, the model must choose the correct class among several possible classes, which increases the chance of error, especially if some classes are underrepresented in the training data. Multi-class classification adds another layer of complexity since each sample may belong to multiple classes simultaneously, requiring the model to understand and predict all applicable classes.

- Imbalance between categories: Often, in multi-category settings, some categories contain far more examples than others. This imbalance can lead to models that are biased towards more frequent categories, reducing their overall effectiveness across less frequent categories. Similarly, in multi-class settings, some classes may be repeated more frequently than others, which can skew the model's predictions.

- Fine-tuning and adaptation: While models like AraBERTv0.2, CAMeLBERT, and ARBERTv2 are pre-trained on a large set of models, their performance on specific tasks such as multi-class classification or multi-label classification can depend on how well they are tuned. Fine-tuning on a task-specific dataset is critical, but without sufficient task-specific data or proper organization, models can overadapt to the training data and perform poorly on unseen data.

- Linguistic nuances: Arabic language processing poses unique challenges due to the richness of the Arabic language in terms of linguistic form and the presence of many homographs (words that are spelled the same way but have different meanings). Effective processing of these nuances requires either specialized pre-processing or structures designed to better capture these aspects, which can be a limitation of general-purpose models such as AraBERTv0.2, CAMeLBERT, and ARBERTv2 when they are not modified for such details.

## 5 Conclusion

In this paper, we conducted a comparative study of deep learning models for a classification task using the Arabic Language Learners' Corpus (ALC). We evaluated three models based on the BERT architecture: CAMeLBERT, AraBERTv0.2, and ARBERTv2. These models were fine-tuned and tested on two classification scenarios: binary and multi-class.
The experimental results indicate that all three models perform exceptionally well in binary classification, with F1 scores of 96.82% for CAMeLBERT, 96.21% for AraBERTv0.2, and 94.95% for ARBERTv2. However, the performance decreased for multi-class classification. CAMeLBERT achieved the highest performance in both subcategories: 60.43% for categorization based on mother tongue and 66.13% for that based on school level. In contrast, the F1 score related to the other two models did not exceed 37% for the first subcategory and 54% for the second one.

The noticeable decrease in performance of the three models in the multi-class classification task can be attributed to two main factors: firstly, the size of the corpus used and the disparities in the number of texts between languages and grade levels; and secondly, the increased complexity of these classifications compared to binary classification.
Comparing the three models, we found that CAMeLBERT's outperforms ARBERTv2 and AraBERTv0.2. This can be ascribed to several parameters: an extensive and diverse training corpus, effective fine-tuning of tasks, architectural innovations, and robust benchmark results.
For future work, we plan to incorporate additional linguistic features such as syntactic and Part of speech tagging to enhance the models' efficiency.

## References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Abdullah Alfaifi, Eric Atwell, and Hedaya Ibraheem. 2014. Arabic learner corpus (alc) v2: a new written and spoken corpus of arabic learners. In *International Symposium Learner Corpus Studies in Asia and the World (LCSAW)*, volume 2, pages 77 – 89. Kobe International Communication Center.

Wissam Antoun, Fady Baly, and Hazem M. Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *CoRR*, abs/2003.00104.

Daniel Blanchard, Joel R. Tetreault, Derrick Higgins, A. Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013:15.

Lingzhen Chen, Carlo Strapparava, and Vivi Nastase. 2017. Improving native language identification by using spelling errors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 542–546, Vancouver, Canada. Association for Computational Linguistics.

Iria del Río. 2020. Native language identification on l2 portuguese. In *Computational Processing of the Portuguese Language*, pages 87–97, Cham. Springer International Publishing.

Iria del Río Gayo, Marcos Zampieri, and Shervin Malmasi. 2018. A Portuguese native language identification dataset. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 291–296, New Orleans, Louisiana. Association for Computational Linguistics.

Meng Joo Er, Rajasekar Venkatesan, and Ning Wang. 2016. An online universal classifier for binary, multi-class and multi-label classification. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 003701–003706.

Dominique Estival, Tanja Gaustad, Son Bao Pham, and Will Radford. 2007. Profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, volume 263, page 272.

John Fields, Kevin Chovanec, and Praveen Madiraju. 2024. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, 12:6518–6531.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with TF-IDF weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 216–223, Atlanta, Georgia. Association for Computational Linguistics.

Anna Golynskaia. 2022. An error coding system for the turkish learner corpus. *The Journal of Linguistics*, 0(39):67–87.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. 2009. *International corpus of learner English*, volume 2. Presses universitaires de Louvain Louvain-la-Neuve.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *CoRR*, abs/2103.06678.

Radu Tudor Ionescu. 2015. A fast algorithm for local rank distance: Application to arabic native language identification. In *Neural Information Processing: 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings, Part II 22*, pages 390–400. Springer.

Scott Jarvis and Magali Paquot. 2015. *Learner corpora and native language identification*, Cambridge Handbooks in Language and Linguistics, page 605–628. Cambridge University Press.

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 the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 382–389, Copenhagen, Denmark. Association for Computational Linguistics.

Kristopher Kyle, Scott Andrew Crossley, and Youjin Kim. 2015. Native language identification and writing proficiency.

Batia Laufer and Nany Girsai. 2008. Form-focused Instruction in Second Language Vocabulary Learning: A Case for Contrastive Analysis and Translation. *Applied Linguistics*, 29(4):694–716.

Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. A deep generative approach to native language identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1778–1783, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shervin Malmasi, Iria del Río, and Marcos Zampieri. 2018. Portuguese native language identification. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 115–124. Springer.

Shervin Malmasi and Mark Dras. 2014. Arabic native language identification. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 180–186, Doha, Qatar. Association for Computational Linguistics.

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, Mark Dras, and Irina Temnikova. 2015. Norwegian native language identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 404–412, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Seifeddine Mechti, Nabil Khoufi, and Lamia Hadrich Belguith. 2020. Improving native language identification model with syntactic features: Case of arabic. In *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 2*, pages 202–211. Springer.

Elham Mohammadi, Hadi Veisi, and Hessam Amini. 2017. Native language identification using a mixture of character and word n-grams. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 210–216, Copenhagen, Denmark. Association for Computational Linguistics.

Yao Qian, Keelan Evanini, Xinhao Wang, David Suendermann-Oeft, Robert A. Pugh, Patrick L. Lange, Hillary R. Molloy, and Frank K. Soong. 2017. Improving Sub-Phone Modeling for Better Native Language Identification with Non-Native English Speech. In *Proc. Interspeech 2017*, pages 2586–2590.

Nikita Remnev. 2019. Native language identification for russian. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 1–7.

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

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 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*, pages 48–57, Atlanta, Georgia. Association for Computational Linguistics.

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 2012*, pages 2585–2602, Mumbai, India. The COLING 2012 Organizing Committee.

Ahmet Yavuz Uluslu. 2023. Turkish native language identification. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 303–307, Online. Association for Computational Linguistics.

Ahmet Yavuz Uluslu and Gerold Schneider. 2022. Scaling native language identification with transformer adapters. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 298–302, Trento, Italy. Association for Computational Linguistics.

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

Marcos Zampieri, Alina Maria Ciobanu, and Liviu P Dinu. 2017. Native language identification on text and speech. *arXiv preprint arXiv:1707.07182*.