Turkish Native Language Identification

Ahmet Yavuz Uluslu
Universität Zürich & PRODAFT
ahmetyavuz.uluslu@uzh.ch

Gerold Schneider
Universität Zürich
gschneid@cl.uzh.ch

Abstract

In this paper, we present the first application of Native Language Identification (NLI) for the Turkish language. NLI involves predicting the author's native language by analysing their writing in other languages. While most NLI research has focused on English, our study extends its scope to Turkish. We used the recently constructed Turkish Learner Corpus and employed a combination of three syntactic features (CFG production rules, part-of-speech n-grams and function words) with L2 texts to demonstrate their effectiveness in this task.

1 Introduction

Native Language Identification (NLI) is the task of automatically identifying the native language (L1) of an individual based on their linguistic productions in another language (L2). The underlying hypothesis is that the L1 influences learners' second language writing as a result of the language transfer effect (Yu and Odlin, 2016). It is used for a variety of purposes including forensics applications in cybercrime (Perkins, 2018) and secondary language acquisition (Swanson and Charniak, 2014).

Research in NLI is mainly conducted with learner corpora, which comprise collections of writings by individuals learning a new language. These writings are annotated with metadata such as the author's native language (L1) or their fluency level. Recent NLI studies on languages other than English include Portuguese (del Río Gayo et al., 2018), Arabic (Malmasi and Dras, 2014a), and Chinese (Malmasi and Dras, 2014b). The learner corpus is the backbone of NLI research, which means that extending research to a novel language depends on acquiring the appropriate learner corpora for that language. In the past, studies have focused on L2 English because of the prominence of this language in language research and the relatively large amount of data available. To the best of our knowledge, this study presents the first detailed NLI experiments on L2 Turkish. We employ the recently constructed Turkish Learner Corpus (TLC) (Anna, 2022) and investigate widely used linguistic features for NLI. The remainder of the paper is organised as follows: Section 2 discusses related work in NLI, Section 3 and 4 describes the methodology and dataset used in our experiments, and Section 5 presents the experimental results. Finally, Section 6 presents a brief discussion and concludes this paper with directions for further research.

2 Related Work

NLI is typically modeled as a supervised multiclass classification task. In this experimental design, the individual writings of learners are used to train a model while the author's L1 information serves as class labels. A variety of feature types at the syntactic and lexical levels were studied to capture distinct characteristics of the language interference phenomenon: spelling errors, word and lemma n-grams, dependency parsing, and morphosyntax. A more detailed review of feature extraction-based methods can be found in two shared task reports on the NLI task organised in 2013 and 2017 (Tetreault et al., 2013; Malmasi et al., 2017).

In recent years, there has been increased experimentation with deep learning methods, including pre-trained transformers (Steinbakken and Gambäck, 2020) and generative models (**?**). While these models slightly outperformed the state-of-the-art performance achieved by feature-based stacked classifiers, questions about their interpretability, inherent biases, and practical shortcomings in industrial applications remain unexplored. Traditional methods based on hand-crafted features continue to be preferred in many implementations due to their simplicity in training and resource efficiency. Within this context, Uluslu and Schneider (2022) approached the NLI scalability problem in the context of cybercrime through the use of adapter fine-

tuning.

3 Data

In this study, we use data from the TLC (Anna, 2022). TLC is a learner corpus composed of the writings of learners of Turkish. These texts are essays written as part of a test of Turkish as a secondary language. Each text includes additional metadata such as the nationality of the author and the genre of the text. The corpus also includes error codes and corrections, although we do not make use of this information.

We used a subset of the dataset containing texts for five L1 groups: Arabic (ARA), Albanian (AL), Azeri Turkish (AZ), Farsi (IR), and Afghani (Pashto) (AFG). We chose these five languages mainly because of the immigration trend observed in Turkey, which will result in a need for additional capabilities for cybercrime forensics and language learning applications for educational institutions. We limit our study to the genre of essays. The other genres (letters of different forms) in the corpus are unbalanced and scarce which may introduce linguistic biases across different registers. We also do not attempt to adjust the dataset based on the writing prompts because they were unbalanced across languages. We randomly selected sentences from the same L1 and combined them to produce documents of approximately the same length. This methodology ensures that the texts for each L1 are a mix of different authoring styles, topics, and language proficiency. The composition of our data is shown in Table 1.

L1	Docs	Tokens	TTR	Avg Words
AFG	55	12546	0.47	278.8
AL	58	15001	0.52	250.6
ARA	65	16969	0.49	252.9
AZ	54	15850	0.47	273.5
IR	52	12870	0.51	246.4

Table 1: Distribution of the five L1s in terms of texts, tokens, type/token ratio (TTR) and average words.

4 Methodology

4.1 Classifier

In our study, we use the standard supervised multiclass classification approach for NLI. A linear Support Vector Machine (SVM) is used for classification and feature vectors are created using a TF-IDF

weighting scheme, in line with previous research (Gebre et al., 2013). We initially experimented with relative frequencies but obtained better preliminary results with TF-IDF. We performed a grid search in parameter space for the regularisation parameter C in the range 10e-6 to 10e-1 and set max_iteration = 5k to ensure model convergence. We find that the generalisation of the model reaches its limit at C = 1, we, therefore, choose this value.

4.2 Evaluation

Following the previous NLI studies, we present our findings using classification accuracy through 10-fold cross-validation (10FCV), which has become the standard for NLI result reporting in recent years. Our cross-validation approach is randomised and stratified, aiming to maintain consistent class proportions across partitions. Since our dataset is slightly unbalanced, we provide detailed metrics in addition to accuracy, including precision per class, recall, and F1 values. We also compare these results to a random baseline.

4.3 Linguistic Features

We focus only on content-independent features, in particular syntactic features, following the example of studies on NLI in other languages (Malmasi et al., 2015). Due to the imbalance in the topic distribution in the TLC corpus, we decided not to include lexical features such as word n-grams and embeddings in our study. Topic bias can arise when certain subjects or topics are not equally represented across different classes (Brooke and Hirst, 2013). For example, only students with Azeri and Farsi L1 were asked to respond to prompts specifically about happiness and time. This can result in the classifier to associate these topics with the languages, rather than discerning the linguistic characteristics inherent to Azeri and Farsi, thereby introducing a confounding variable to the task. Even if we attempted to balance the topics across languages, similar to the TOEFL11 dataset (Blanchard et al., 2013), we found that particular rhetoric strongly influences certain backgrounds. For example, writings of the students from Afghanistan were predominantly religious, regardless of the topic. By focussing on syntactic features, we aim to capture the underlying syntactic influence of the L1 on its L2 writing independently of the content. We experimented using a combination of three syntactic features: context-free grammar (CFG) production rules, part-of-speech n-grams, and function words.

Function words are content-independent words, including prepositions, articles, and auxiliary verbs, that play a crucial role in conveying grammatical relationships between words. It is often challenging for L2 speakers to use the appropriate function words and production errors may be due to the influence of their L1 (Schneider and Gilquin, 2016). These function words are recognised as valuable features for the NLI task. We extracted frequencies of 75 Turkish words from different grammatical categories. However, it's worth noting that many grammatical aspects, which are morphologically expressed in Turkish, may not be as strongly captured as they would be in English.

Part-of-Speech tags are linguistic categories or word classes that signify the syntactic role of each word in a sentence. They include basic categories such as verbs, nouns, and adjectives. Assigning POS tags to words in a text introduces a level of linguistic abstraction, meaning that we can work with the underlying structure rather than the content. We use the Turkish POS module from Stanza (Qi et al., 2020) to extract universal POS tags, from which we create n-grams of sizes 1 to 3. These n-grams serve to capture preferences for specific word classes and their localised ordering patterns. Our experiments indicated that sequences of order 4 or higher lead to lower accuracy due to the limited size of our corpus. Therefore, we excluded such higher-order n-grams from our analysis.

Hızlı	*kahverengi*	*tilki*	*ve*
ADJ	ADJ	NOUN	CONJ
tembel	*köpeğin*	*üzerinden*	*atlar.*
ADJ	NOUN	POSTP	VERB

3-gram Example: (NOUN, POSTP, VERB)
Functional n-gram Example: (ve, üzerinden)
CFG Rule example: (NP → ADJ NOUN)

Figure 1: An example of a Turkish sentence and feature extractions for POS n-grams, function word n-grams and CFG-Rule extractions.

CFG production rules are used to generate constituent parts of sentences, such as noun and verb phrases. We use the Turkish parsing module of Stanza (Qi et al., 2020) to extract the constituency tree for the documents. The production rules are then extracted and each rule is used as a standalone feature. We exclude lexicalizations to focus on more abstract and general syntactic patterns. These production rules can encode highly idiosyncratic constructions that are specific to particular L1 groups. They have been widely utilized in various ensemble methods for NLI and have been shown to complement other features effectively (Malmasi and Dras, 2018).

5 Results

In this section, we present the results in terms of accuracy achieved by individual feature types. Subsequently, we report the performance obtained using the combination of all features. Finally, we examine the performance obtained by the best system for each L1 class.

Feature Type	Accuracy (%)
Random Baseline	20.0
POS 1-grams	33.4
POS 2-grams	38.9
POS 3-grams	38.6
Function Words	37.2
CFG Production Rules	41.4
Full Combination	**44.2**

Table 2: 10-FCV Accuracy Classification Results

Table 2 displays the results of the systems trained with different feature types in terms of accuracy. We found that all feature types individually outperform the baseline. The CFG rules are the features that individually perform the best, achieving an accuracy of 41.4%. This demonstrates the importance of the syntactic differences between the L1 groups. The full combination, using all feature types, obtains performance higher than CFG features achieving 44.2% accuracy. These trends are very similar to previous research using the same features (del Río Gayo et al., 2018; Malmasi and Dras, 2014a) with comparable corpora.

L1	Precision	Recall	F1-score
AFG	0.50	0.29	0.37
AL	0.45	0.54	0.49
ARA	0.43	0.65	0.52
AZ	0.47	0.41	0.44
IR	0.37	0.15	0.21
Average	0.44	0.41	0.40

Table 3: Full combination per-class results: precision, recall and the F1-score.

Table 3 shows the results obtained for each L1 in terms of precision, recall, and F1 score, as well as the average results for the five classes. Across all classes, we obtain a micro-averaged F1 score of 0.40 and a macro-averaged F1 score of 0.44.

To provide a visual representation of these findings and to highlight any error patterns, we present a heatmap confusion matrix of the classification errors in Figure 2.
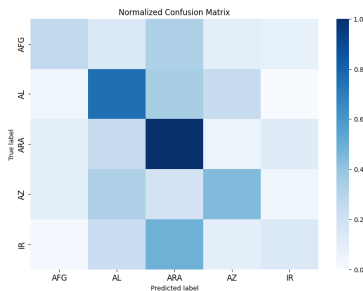

Figure 2: Confusion Matrix of Classification Errors

In most NLI studies, the expected difficulty is to distinguish closely related L1s that may belong to the same language family. Based on the analysis of the confusion matrix, the most notable confusion occurs between Persian and Arabic, both of which have strong lexical connections to the Turkish language. However, since we are working with content-independent features, we attribute this confusion to corpus representation and do not seek any linguistic explanations. We observed no confusion between Afghani and Persian, even though both languages belong to the same language family and have strong similarities. A previous study in comparable settings also failed to offer strong interpretations based on sociolinguistic insights in the error analysis (Malmasi and Dras, 2014a).

Our analysis brings attention to two potential limitations that might prevent drawing connections between model errors. Firstly, although the size of our corpora is relatively limited when compared to other NLI studies—being five times smaller than the Portuguese corpus reported by del Río Gayo et al. (2018) and ten times smaller than the Norwegian corpus described in Malmasi et al. (2015)—it is comparable to the corpus size used for Arabic (Malmasi and Dras, 2014a), which achieved a similar performance compared to our study. The difference in data size and quality might explain the model's generalisation capabilities. Secondly, we acknowledge that our parser might not be entirely suitable for learner language, which could

introduce additional noise into the feature space (Van Rooy and Schäfer, 2009).

6 Conclusion & Discussion

In this study, we presented the first experiments with Turkish NLI and achieved a level of performance comparable to previous results for other languages. Our main focus was to investigate the effectiveness of syntactic features for Turkish, a language that differs from English in certain aspects, particularly in morphological complexity. Another significant contribution of our work is the introduction of a new dataset for NLI, specifically designed to address L1-based language transfer effects. This corpus can serve as a valuable resource for researchers to validate and refine their methodologies across various datasets and languages.

We identify several promising directions for future research. Firstly, we plan to expand the corpus by incorporating more learner writings and extending the analysis to encompass other L1 languages. Additionally, we believe that assessing the proficiency level of learners can shed further light on the observed challenges. Finally, we plan to explore more linguistically sophisticated features in our investigation. For instance, leveraging L1 mistakes from a morphological perspective as a content-independent feature could yield valuable insights. To this end, our follow-up study will incorporate a broader range of features to enhance the robustness and comprehensiveness of our analysis.

Ethics Statement

Acknowledgements

References

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

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Julian Brooke and Graeme Hirst. 2013. Native language detection with 'cheap'learner corpora. In *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)*, volume 1, page 37.

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.

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.

Shervin Malmasi and Mark Dras. 2014a. 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. 2014b. Chinese native language identification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 95–99, Gothenburg, Sweden. Association for Computational Linguistics.

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

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.

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 the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark. Association for Computational Linguistics.

Ria Perkins. 2018. The application of forensic linguistics in cybercrime investigations. *Policing: A Journal of Policy and Practice*, 15.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Gerold Schneider and Gaëtanelle Gilquin. 2016. Detecting innovations in a parsed corpus of learner english. *International Journal of Learner Corpus Research*, 2:177–204.

Stian Steinbakken and Björn Gambäck. 2020. Native-language identification with attention. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 261–271, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Ben Swanson and Eugene Charniak. 2014. Data driven language transfer hypotheses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 169–173, Gothenburg, Sweden. Association for Computational Linguistics.

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.

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.

Bertus Van Rooy and Lande Schäfer. 2009. The effect of learner errors on pos tag errors during automatic pos tagging. *Southern African Linguistics and Applied Language Studies*, 20:325–335.

Liming Yu and Terence Odlin. 2016. *New perspectives on transfer in second language learning*, volume 92. Multilingual Matters.