# Multiclass Native Language Identification and Language Proficiency Identification: Feature Analysis on Toefl-11 Dataset

Robin De Paepe
University of Antwerp

*This paper presents the results of implementing a Linear Support Vector (SVM) Machine Learning approach to the TOEFL-11 dataset. For this Shared task, we've attempted to both perform Native Language Identification(L1) and Language Proficiency testing (L2) on the dataset. The model used in during this task was constructed on the basis on some very prominent NLP methods. The model performed reasonably well on the Native Language Identification section of the task with an overqll accuracy score of 81%. L2 Language proficiency predictions lagged somewhat behind at 67% macro f1-score. The model proposed made use of character, word and part-of-speech tags in combination with several syntactic features.*

## 1. Introduction

The continued growth of the field of NLP within the digital Humanities has seen been offset by a parallel rise in popularity in NLP machine learning techniques. Native Language Identification, or NLI, benefits especially from the pioneering insights in the field of NLP. Native-language identification (NLI) is the task of identifying an author's first language (L1) given only information expressed in the author's second language (L2). NLI has been a popular tool for educational purposes such as developing grammatical error correction systems which can personalize their feedback and model performance to the native language of the user.(Rozovskaya and Roth 2011) Additionally, NLI has also seen widespread usage in authorship identification, forensic analysis, multi-author text attribution and Second Language Acquisition Research. (Malmasi et al. 2017)

Language Proficiency Identification on the other hand tackles the classification of different proficiency levels within a certain spoken or written language, most often as an author's second language (L2). Language proficiency is very ubiquitous within almost any professional field since Second language proficiency level indicates how well a person comprehends and communicates in a language other than the person's heritage language. Having higher language proficiency implies that the person is able to use the language competently while lower language proficiency implies otherwise. (Blanchard et al. 2013) Where NLP and machine learning techniques prove especially useful in this field is in the potential to create automated models which can asses language proficiency levels of authors in professional environments to a much more efficient extent in comparison to human-laboured proficiency testing. This is due to the fact that proficiency testing has to be done in accordance to standardized set of testing executed by professionals. As shown by Settles, T. LaFlair, and Hagiwara (2020), this can prove cumbersome due to the many of the procedures and requirements for planning, creating, revising, administering, analyzing, and reporting on high-stakes tests and their development.

Throughout this paper we will display experiments done in the context of the Shared Task at the University of Antwerp among students of the Masters in Digital Text Analysis. The Shared Task was conducted on the well-known TOEFL-11 data-set in which we attempted to provide a machine learning model that is able to predict the native language(L1) of a speaker and the proficiency in English as a second language (L2) from English-written essays provided by the data set. First we will asses (in Section 2) recent state of the art models and techniques that have been used in the academic field as a theoretical framework for the experiments. In section 3 we will asses the specifics of the data-set used during the Task (3.1), after which we will explain the choose of

methods which were used (3.2), after which we will dive into the specifics of the feature selection (3.3). Finally in section 3.4 we will evaluate the results of the experiments themselves. Section 4 contain the discussion of our experiments in the wider perspective of the theoretical framework provided in section 2. Finally section 6 summarized the findings of this paper in the light of future potential research.

## 2. Related Research

Previous work on native language identification has already shown that a wide array of features can improve performance on L1 language detection systems. Koppel, Schler, and Zigdon (2005) showed that the use of character n-grams, function words and part-of-speech bi-grams, can result in vast improvements during automatic detection. Previous work by Wong and Dras (2011) also has showed the benefits of implementing syntactic structures like function words in a NLI detection model, however, such features will not be explored in this research here. As Yang, Yu, and Lim (2016) states in their work on L2 proficiency identification show, prior research like that of Crossley, Salsbury, and McNamara (2012), makes wide use of NLP methods and techniques, most notably those used for NLI in particular. However, a wide range of these explorations seem to mainly focus on the implementation of spoken language tests as data. (van der Walt, de Wet, and Niesler 2008; Luo et al. 2008) Tsur and Rappoport (2007) reported that choice of words in second language writing is highly influenced by the frequency of native languages syllables which can translate into errors in L2 writing. Frequency counts and word based n-grams where also used in a very accurate experiment on the Europarl corpus, where and accuracy between 87% and97% was achieved. Unfortunately, these high accuracy scores are mainly obtained through the identification of very particular phrases used in parliamentary contexts. (Van Halteren 2008)

The methodology of this research mainly builds on the concept of error analysis and error detection as initially conceptualized by Corder (1967). Corder showed that 'errors made in second language learning provide evidence that a learner uses a definite system of language at every point in his development'. Agarwal, Agarwal, and Mittal (2014) already showed a very practical implementation of error analysis implementation methodology in the creation of a web-based error analysis tool with the use for computational linguists and NLI- implementations in mind.

Additionally, this paper also builds forth on the concepts of contrastive analysis, which suggests that errors made by second language users, originate from habit formation created during learning of the native language.The Contrastive Analysis Hypothesis (CAH) formulated by Robert Lado, argues that errors potentially made by learners of a second language are predicted from interference by the native language. Such a phenomenon is usually known as negative transfer. (Lado 1957) However, as argued by Wardhaugh (1970) in his differentiation between strong and weak forms of contrastive analysis, this paper follows the trial of Wong and Dras (2011) in exploring the weak form of CAH. This weak form proposes that potential differences in syntactic transfer cannot form the sole explanation for learning difficulties in second languages as opposed to the strong form, where all errors made are attributed to the native language.

This paper also notes that recently there has been a shift towards using performance on cognitive tasks as features within automated NLI task like illustrated in Yang, Yu, and Lim (2016). Yang illustrates through the use lexical decision tasks (LDT), competitive accuracy in L2 language proficiency can be obtained. Although these methods do look promising and invite further exploration. The method proposed by Yang seems more suited the absence of the availability of conventional language proficiency test

like the TOEFL-11 data-set due to the relative lower efficiency costs the collection of LDT requires compared to TOEFL corpuses for example.

## 3. Experiment description

### 3.1 Data

For this study, we made use of the TOEFL11 data set, which consists of a Corpus of non-native written English essays with authors from 11 different languages.This ETS Corpus is specifically targeted toward NLP research, for native language identification in particular.(Blanchard et al. 2013) The essays are evenly distributed among authors with 11 different natives speaking languages(Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu and Turkish), which are provided in the training set. Additionally the 3-level proficiency levels (high, medium, low) are provided for each essay. As we can see from figure 1, the distribution of proficiency levels was not generated evenly in contrast to the the native languages.

The TOEFL11 corpus was designed specifically to support the task of native language identification. Because all of the essays were collected through ETS's operational test delivery system for the TOEFL-test, the encoding and storage of all texts in the corpus is consistent.(Tetreault, Blanchard, and Cahill 2013) Our test data consists of 1.100 essays of which the native language(L1) and language proficiency levels were unknown during the experiments. What makes the TOEFL corpus especially useful optimal for our experiments is the fact that it contains a large collection of training data to train our systems on. CITATION NEEDED

**Table 1**

| Number of essays (train): 11000<br>train: Language | |
|---|---|
| ARA | 1000 |
| DEU | 1000 |
| FRA | 1000 |
| HIN | 1000 |
| ITA | 1000 |
| JPN | 1000 |
| KOR | 1000 |
| SPA | 1000 |
| TEL | 1000 |
| TUR | 1000 |
| ZHO | 1000 |
| train: Proficiency | |
| high | 3835 |
| low | 1201 |
| medium | 5964 |

### 3.2 Methods

Our model incorporates an array of textual and numerical features, i.e. word n-grams, character n-grams, POS n-grams, as well as several numerical features extracted from the data-set like proportion of capitalized words, punctuation and average sentence length. System performance was measured in terms of both classification accuracy on a development set (20 percent of training data) as well as under 10-fold cross-validation. For the NLI task, system performance was measured by accuracy, while we measured performance on the language proficiency task by F1-score. Total perfomance on the model was based on the average of both scores. Our system made use of a Linear Support Vector Machine (SVM) classification model to train on our textual and numerical features we provided. Textual features were extracted through term-frequency-inverse document frequency (tf-idf) weighting. The methods applied during the experiments

were based on the methods and theoretical framework which are described in Section 2.

### 3.3 Feature selection

We used the same model for both task performed, however, during experimentation, we performed extensive feature simplification testing and found the especially our Language Proficiency predictions did not seem to benefit from several features which were implemented during the NLI task.(see section 3.4) As mention in our theoretical framework, NLI in particular benefits from introducing word and character features into the model. (Jarvis, Bestgen, and Pepper 2013) Additionally several other features which focused primarily on punctuation and capitalisation within the written text, were implemented to complement our prior features. This model made use of proportion of capitalized words, proportion of capitalized 'i"s, proportion of punctuation and average sentence length of each essay.

**3.3.1 Word, character and POS n-grams.** lexical authenticity of an author can be extracted through word and character features which can help in identifying specific patterns among different native language authors. Additionally as Jarvis, Bestgen, and Pepper (2013); Malmasi and Dras (2015) have shown, POS features are able to capture the morpho-syntactic patterns which are apparent in each text. word an POS n-grams were used ranged between with n ranging between 1 and 2. For POS n-grams we use a range of 1 to 3 n-grams and character n-grams were used within the range of 2 to 4. POS and n-grams were all extracted from the data-set using the spaCy software package. CITATION?

Since the topics in the essays throughout the TOEFL11 data-set are predefined and distributed evenly among the data-set, the limited usefulness of word n-grams due to

the fact that these are often topic-bound as described by Kochmar (2011) are relatively negated. Additionally, it has been shown that word n-grams can reveal specific errors related to the native language of the speaker. (Chen, Strapparava, and Nastase 2017)

As shown in a number of works previously performing NLI on similar data-set, it has been shown that character n-grams perform particularly well on classifying an author's L1 since it reflects spelling errors and sound-to-character mappings particularly. (Koppel, Schler, and Zigdon 2005) Works like those of Wong and Dras (2011) also show that character n-grams are particularly useful in displaying the influence in phonology of the native language.

POS n-grams have also proven to be useful for NLI and Proficiency in a number of prior works as well as they are able to translate grammatical properties and word orders of the author into the model. They can also 'capture idiosyncratic constructions and sequences of words not typical for English'(Kochmar 2011)

**3.3.2 Punctuation & capitalisation.** The model also made use of additional features which mainly focused on the ability for second language users to implement appropriate placement of punctuation and capitalisation. We first extracted the proportion of punctuation marks in comparison to the total number of POS-tags in each individual essay. Secondly, we went on to calculate the average sentence length within each essay. As shown by Cimino et al. (2018), sentence length can display 'parse trees of similar depth and complement chains among linguistic profiles.'

According to Shatz (2019), Capitalisation is a salient orthographic feature, which plays an important role in linguistic processing during reading, and in writing assessment. Learners' second language (L2) capitalisation skills are influenced by their native language (L1). Because of this, we implemented two features in the model which takes advantage of these distinct differences. We first extracted the proportion of capitalized

words from each essay. Secondly we extracted the number of capitalized words at the beginning of each sentence. This distinction is important to make since capitalisation across sentences instead of only in the beginning of each sentence can be an indicator of L1's grammar transfer like with noun capitalisation in German. (Kochmar 2011) Finally we also added a feature which registered the proportion of capitalized 'i's, since this pronoun should always be capitalized in English. It is possible that differences among second language users can be detected this way.

.

**3.4 Experiments**

While deciding on the algorithm to use, simulated several dummy-classifiers on the data and obtained a uniform baseline of 0.087, a most frequent baseline: 0.091 and a stratified baseline: 0.086 on the NLI task while only obtaining baseline scores of 0.34 , 0.54, 0.43 on the proficiency classification task respectively. After confirmation of our dummy classifiers on the data-set, we were able to confirm prior reports on SVM's effectiveness in NLI tasks as this performed as one of the best algorithms for our two text categorization tasks. (Markov et al. 2017) We ultimately decided on going for a Linear Support Vector Machine algorithm while using a TF-IDF vectorizer in our pipeline for our textual data. After initial baseline testing between a regular and linear SVM algorithm in combination with either a CountVectorizer or a TF-IDF vectorizer. Our final combination performed best during the initial testing in particular on the NLI task with an initial accuracy of 0.76 while giving similar result compared to the other combinations on the proficiency identification with an F1-score of 0.58.

After the model selection, extensive feature testing was performed on both tasks. Initial baselines for both tasks were set at an accuracy of 0.79 for the NLI task while the F1-score for the Proficiency task was set at 0.63. The baseline contained all described

features from section 3.3. We've attempted at performing Feature simplification with several numeric features we extracted. Improvements depended highly on the task we were testing for. L1 identification did only benefit from leaving out proportion of capitalisation with an improvement from 0.79 towards 0.81 accuracy. Proficiency classification on the other hand benefited considerably more from feature simplification: When leaving out both average sentence length and proportion of capitalisation. baseline f1-score improved from 0.63 towards 0.66.

Finally, we performed grid-search optimization on our model for both tasks independently. Parameter optimization was performed on word n-grams, character n-grams and the POS n-gram ranges. Parameter optimization did not generate significant improvements on either task. Ultimately grid-search optimization made us settle ranges from bi-grams to 4-grams in our character pipeline, between uni-grams and tri-grams on our word n-grams and between one-grams and bi-grams for the POS-tags for the NLI task. For the Proficiency task, we applied ranges between bi-grams and tri-grams for characters, between uni-grams and bi-grams for word n-grams and finally uni-grams and 4-grams for POS-tags.

**3.4.1 Evaluation & Results.** The model's performance was tested on a development set which consisted of a 10% split of the training data. The following table displays the performance of the model on the NLI task through a classification report on the development set.

The classification report obtained an overall accuracy of 0.81 on the NLI task, the best performing classes consisted of German, Italian, French and Chinese while the model seemed to perform least well on languages like Telugu and Hindi. Figure 1 shows that the model seems to have particular difficulties with distinguishing Hindu and Telulu authors from each other, as these two languages are often mixed up for each

**Table 2**
L1 Identification 10-fold cross-validation

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| ARA | 0.84 | 0.75 | 0.79 | 1000 |
| DEU | 0.86 | 0.93 | 0.89 | 1000 |
| FRA | 0.83 | 0.82 | 0.83 | 1000 |
| HIN | 0.70 | 0.74 | 0.72 | 1000 |
| ITA | 0.87 | 0.86 | 0.87 | 1000 |
| JPN | 0.82 | 0.79 | 0.81 | 1000 |
| KOR | 0.78 | 0.78 | 0.78 | 1000 |
| SPA | 0.76 | 0.77 | 0.77 | 1000 |
| TEL | 0.78 | 0.77 | 0.77 | 1000 |
| TUR | 0.86 | 0.86 | 0.86 | 1000 |
| ZHO | 0.84 | 0.87 | 0.86 | 1000 |
| | | | | |
| accuracy | | | 0.81 | 11000 |
| macro avg | 0.81 | 0.81 | 0.81 | 11000 |
| weighted avg | 0.81 | 0.81 | 0.81 | 11000 |

other. Japanese, the second most misidentified language has particular trouble with getting differentiated from both Chinese and Korean, although both these languages obtained relatively good accuracies themselves. (0.82 and 0.88) respectively. Two other noteworthy findings in the NLI results is the relative 'general' high misclassification of Arabian authors. In contrast to previously mentioned languages Arabian L1 speakers seem to have higher false negative labels among several languages as opposed to just one or two like with Japanese. Finally, French speakers also seem to get relatively often misidentified as German or Italian L1 users by the model.

Although lower in accuracy, these findings seem to be in line with prior works on similar data-sets, where language identification systems performed best on German and Chinese author profiles and had similar troubles with languages like Hindu  Telulu and Japanese and Korean. (Markov et al. 2017; Tetreault, Blanchard, and Cahill 2013) Next, we performed the 10-fold cross-validation method on the total training data-set.

**Table 3**
10-fold cross validation on L2 Proficiency levels

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| high | 0.66 | 0.79 | 0.72 | 3835 |
| low | 0.81 | 0.34 | 0.48 | 1201 |
| medium | 0.73 | 0.72 | 0.73 | 5964 |
| | | | | |
| accuracy | | | 0.71 | 11000 |
| macro avg | 0.73 | 0.62 | 0.64 | 11000 |
| weighted avg | 0.71 | 0.71 | 0.70 | 11000 |

The system performed similar to the classification report with accuracy scores ranging between 0.7736 and a best score 0.8327.

Similarly, we performed both a classification report as well as 10-fold cross-validation during the proficiency identification task. The final proficiency model performed considerably lower in comparison to the model in the NLI task. As we can see from Table 3, proficiency identification only reached 0.64 macro F1-score while obtaining 0.71 on accuracy and 0.70 on weighted average. The model seems to have particular trouble with identifying relevant low-proficiency essays, as can be observed in the recall column. Although precision on classifying low-proficiency texts seems to be relatively high compared to high and medium proficiency classification, the proposed system seems to have particular trouble with retrieving relevant instances for identification. A possible explanation for this could be due to the fact that (see Table 1) the amount training data for low-proficiency text is significantly lower compared to high, and especially medium-proficiency texts. This is also supported by the fact that precision and recall scores are much more balanced with the latter two labels.

## 4. Conclusion

This paper attempted to merge recent methods applied in native language identification tasks with similar works on language proficiency identification tasks. Although results were not as accurate as the experiments where this paper draws inspiration from, we do argue here that the results presented can give new insights in further developing automated systems which seek to combine tasks of NLI and Language Proficiency Identification at the same time. This paper suggests that automated models for both task can apply several overlapping features like character, word and POS n-grams as well several other features like proportion of punctuation used throughout used text as well as capitalisation features. However, the model did not succeed in finding a 'one-fits-all' feature and model selection blueprint for the described texts. Future research in this might further want to explore the relationship in syntactic errors in L2 writing when assessing both L1 identification and L2 proficiency levels. Another aspect which we might want to explore in the future of NLP research is how both tasks perform in multi modal prediction systems. Finally, it should be noted that due to time and space constraints, this experiment did not manage to test all possible experimentation and validation routes that were possible. As an example of this, cross-validation was only performed during the final iteration of the proposed model. We admit here that it certainly would have been more interesting to perform 10-fold cross validation during feature selection and simplification phases. Additionally further exploration of alternative models (like deep learning models) might have been more extensively explored if not for the limited range of this paper.

## References


Agarwal, Apoorv, Ankit Agarwal, and Deepak Mittal. 2014. An error analysis tool for natural language processing and applied machine learning. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 1–5.

Blanchard, Daniel, 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.

Chen, Lingzhen, 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.

Cimino, Andrea, Felice Dell'Orletta, Dominique Brunato, and Giulia Venturi. 2018. Sentences and documents in native language identification. In *CLiC-it*.

Corder, Stephen Pit. 1967. The significance of learner's errors.

Crossley, Scott A, Tom Salsbury, and Danielle S McNamara. 2012. Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2):243–263.

Jarvis, Scott, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118.

Kochmar, Ekaterina. 2011. *Identification of a writer's native language by error analysis*. Ph.D. thesis, Master's thesis, University of Cambridge.

Koppel, Moshe, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *International Conference on Intelligence and Security Informatics*, pages 209–217, Springer.

Lado, Robert. 1957. Sentence structure. *College Composition and Communication*, 8(1):12–16.

Luo, Dean, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose. 2008. Automatic assessment of language proficiency through shadowing. In *2008 6th International Symposium on Chinese Spoken Language Processing*, pages 1–4, IEEE.

Malmasi, Shervin and Mark Dras. 2015. Language identification using classifier ensembles. In *Proceedings of the joint workshop on language technology for closely related languages, varieties and dialects*, pages 35–43.

Malmasi, Shervin, 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, Association for Computational Linguistics, Copenhagen, Denmark.

Markov, Ilia, Lingzhen Chen, Carlo Strapparava, and Grigori Sidorov. 2017. Cic-fbk approach to native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 374–381.

Rozovskaya, Alla and Dan Roth. 2011. Algorithm selection and model adaptation for esl correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 924–933.

Settles, Burr, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, 8:247–263.

Shatz, Itamar. 2019. How native language and l2 proficiency affect efl learners' capitalisation abilities: a large-scale corpus study. *Corpora*, 14(2):173–202.

Tetreault, Joel, 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.

Tsur, Oren and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16.

Van Halteren, Hans. 2008. Source language markers in europarl translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 937–944.

van der Walt, Christa, Febe de Wet, and Thomas Niesler. 2008. Oral proficiency assessment: the use of automatic speech recognition systems. *Southern African Linguistics and Applied Language Studies*, 26(1):135–146.

Wardhaugh, Ronald. 1970. The contrastive analysis hypothesis. *TESOL quarterly*, pages 123–130.

Wong, Sze-Meng Jojo 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.

Yang, YeongWook, WonHee Yu, and HeuiSeok Lim. 2016. Predicting second language proficiency level using linguistic cognitive task and machine learning techniques. *Wireless Personal Communications*, 86(1):271–285.

## 5. Supplementary material

All used models and documentation which was used to execute this research can be found on a private Github repository of the author. Please contact the author of the article if further exploration in the material is required.

**Figure 1**
Confusion Matrix

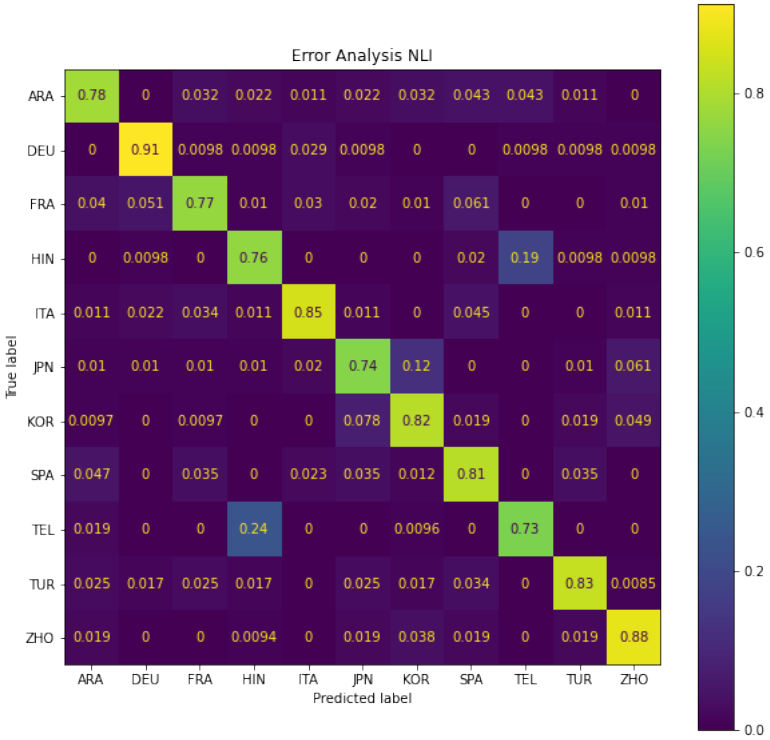

**Figure 2**
Confusion Matrix