# L1ML Task Description

Yuzhe Chen, Jessica Huynh, and Ryan Nicoll

## Introduction

In an increasingly interconnected world, languages spoken on a global scale, such as English, have a growing number of non-native or L2 speakers who demonstrate systematic errors in their English along the adult language acquisition process. Given that their errors tend to be L1-dependent, they can be readily identified, learned, and used in natural language processing applications for native language identification. Native language identification, or NLI, involves classifying the native language of a speaker or writer given linguistic data produced by them in their non-native language. The L1ML team sets out to create an annotation scheme for the ETS Corpus of Non-Native Written English (also known as the TOEFL11) that will identify these errors at various levels of linguistic inquiry (morphological, syntactic, lexical, discourse, etc.) for use in supervised machine learning tasks of native language identification. This will expand upon recent previous work on the corpus that have used methods such as n-gram models based on POS tags, single characters, or function words; Naïve Bayes classifiers; and features based on Stanford dependencies (Tetrault et al., 2013). Our annotated data has the potential to be used in a wide range of applications for NLI, such as security, second-language learning, and grammar correction, among others.

## The Corpus

For our task, we will use the TOEFL11 corpus, which contains 12,100 essays taken from the Test of English as a Foreign Language (i.e., TOEFL), an examination to measure the academic English proficiency for L2 English speakers who wish to enter American universities. The 12,100 essays are answers to eight open-ended prompts and are, on average, about 348 words in length (range: 2 to 876 words).

The corpus contains essays from 11 different languages: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. These languages were chosen for having >1,100 essay samples for use in the corpus. Each essay has been categorized (based on the ETS score of 0 to 5) with a score of Low, Medium, or High, reflecting the level of written English proficiency of the essay.

The L1ML team chose the TOEFL11 corpus because the corpus was assembled specifically with the NLI task in mind, as previous researchers have noted the paucity of cross-linguistic data and difficulty of acquiring additional data. Additionally, the corpus has, to a certain extent, been designed to be balanced: the corpus has taken an evening sampling for each language among each prompt and among each language. However, the corpus has not controlled for score, resulting in an unbalanced distribution of scores per language (see Blanchard et al. 2013: 9 for discussion/statistics). To rectify this, we intend to take an evenly distributed random sampling of each language in order to control for score.

Two observations from the dataset have informed our control method. First, the essays in the High category have few to no observable grammatical errors. The L1ML team had a difficult time distinguishing the essays in this category from those of native speakers. Additionally, the Low and High score sample subsets have numbers skewed towards specific languages: the High Score set has a significantly higher number of German and Hindi essays, while the Low subset has a disproportionate representation of Arabic and Japanese essays. Consequently, we intend to take an even sampling of each language and each prompt from the Medium score set, which will provide enough errors per prompt to be able to provide salient features while still allowing for greater control over the variable of language proficiency in the dataset.

## Annotation Task

For our task, we will be using a tokenized version of the data provided by the assemblers of the TOEFL11 and will use an automatic POS-tagger to assign preliminary parts of speech to each token. Using this pre-processed data, annotators will denote grammatical errors in the following linguistic domains (with included non-exhaustive examples), using a yet-to-be-determined tagset:

- *Morphosyntactic:* article/count errors (**some literatures)*, verb tense/agreement errors (**he walk*, **they eat yesterday*), verb argument errors (**he devoured, *raining often, *she voted at Bernie Sanders),* gender pronoun mismatch (**he rains, *Hillary Clinton and his campaign)*
- *Lexical/semantic:* Orthographic error (**releive*), incorrect use of lexical item (**the circulation* [i.e., traffic] *on Route 25 was very bad)*
- *Discourse:* punctuation inappropriate for discourse (*He, was very lonely)*, awkward/non-native discourse marker (*Bernie Sanders was a Democrat, what's more he was from Vermont)*

This information, once marked up in the data by annotators, can then be used in a machine learning algorithm for a classification task of the L1 of a given speaker given an English L2 text and a known universe of 11 possible languages. The subset of data from our corpus will be further split into training, test, and development subsections for use in this process.

In previous research using the TOEFL11 corpus for NLI, such as the NLI 2013 Shared Task, research teams achieved up to an overall 83% success rate corpus using features such as word, character, and POS n-grams and algorithms such as SVMs and MaxEnt (Tetrault et al, 2013: 54-55). The goal of L1ML is to provide further linguistic information in L2 errors in the hopes of matching or exceeding these results and improving the state of native language identification.

## Works Referenced

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. *HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task.* The 7th Workshop on the Innovative Use of NLP for Building Educational Applications (pp. 54-62). Montreal, QC, CA.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. *TOEFL11: A Corpus of Non-Native English*. Technical report, Educational Testing Service.

Shervin Malmasi and Mark Dras. 2014. *Language Transfer Hypothesis with SVM Weights.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (pp. 1385 - 1390). Doha, Qatar: Association for Computational Linguistics.

Shervin Malmasi, Sze-Meng Jojo Wang, and Mark Dras. 2013. *NLI Shared Task 2013: MQ Submission.* Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 54-62). Montreal, QC, CA.

Joel Tetrault, Daniel Blanchard., and Aoife Cahill. 2013. *A Report on the First Native Language Identification Shared Task.* Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 48–57). Stroudsburg, PA: Association for Computational Linguistics.

Joel Tetrault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. *Native tongues, lost and found: Resources and empirical evaluations in native language identification*. Proceedings of the 24th International Conference on Computational Linguistics (pp. 2585–2602). Stroudsburg, PA: Association of Computational Linguistics.

Sze-Meng Jojo Wang, Mark Dras. 2011. *Exploiting Parse Structures for Native Language Identification.* Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 1600-1610). Edinburgh, Scotland, UK.