Paper#1: A Report on the First Native Language Identification Shared Task

Paper link: [A Report on the First Native Language Identification Shared Task] (Tetreault, Blanchard & Cahill, 2013)

Paper Reading – First Native Language Identification (NLI) Shared Task

This post summarizes the first shared task on *Native Language Identification (NLI)*- predicting a writer's native language (L1) from essays written in a learned language (here, English). It standardizes data, tasks, and evaluation to enable meaningful comparison across 29 participating teams, and remains a foundational benchmark for educational NLP and authorship profiling.

Why this matters

NLI supports **targeted feedback for language learners** (different L1s show distinct error tendencies) and contributes to **authorship profiling**. Before this effort, research relied on small, inconsistent corpora (often ICLE), making results hard to compare. This shared task fixed that by providing a large, balanced corpus and uniform evaluation.

Dataset — TOEFL11

The task introduced the **TOEFL11** corpus: roughly **1,100 essays per L1** across **11 L1s** (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish), sampled across **8 prompts** and proficiency levels (low/medium/high). Splits per L1: **900 train**, **100 dev**, **100 test**. The corpus was curated to minimize topic bias and ensure consistent encoding.

Task setup

Three subtasks evaluated robustness to data availability and domain match:

- **Closed**: Train only on TOEFL11-TRAIN (+ optional DEV).
- **Open-1**: Train on any external data (no TOEFL11); test on TOEFL11-TEST.
- **Open-2**: Train on TOEFL11 + any external data.

Methods & What Worked

Learning algorithms

The overwhelming majority used **Support Vector Machines (SVMs)**, reflecting the high-dimensional, sparse nature of n-gram feature spaces. Other approaches included **MaxEnt/Logistic Regression**, **ensembles**, **string kernels (character n-gram kernels)**, **Discriminant Function Analysis**, **k-NN**, and a few specialized methods (e.g., PPM). Notably, top systems in the closed task were SVM variants, and at least one team (BUC) used *string kernels over character features* effectively.

Feature engineering

Surface features dominated. Teams relied heavily on:

- **Character n-grams** (often up to 5; several top teams used higher orders- up to 7-9).
- **Word n-grams** (typically 1-3; a few tried 4-5).
- **POS n-grams** (1-4, occasionally 5).
- **Function word distributions** and stylistic counts (length of words/sentences, document length).
- **Error and spelling features** (limited adoption, but conceptually relevant for L1 transfer).
- **Syntactic features** (dependency rules, CFG productions, TSG fragments) — tried by several teams; gains were modest compared to lexical/character n-grams.

Preprocessing & representation choices

Common choices included lowercasing, pruning rare n-grams, and using **binary presence vs. normalized counts**. Many teams compared TF, TF-IDF, and binary encodings; binary often worked well with character n-grams. Some experimented with feature selection by top-*k* n-grams versus “use all” within order constraints.

Ensembling & kernels

A subset combined multiple base learners (e.g., SVMs with different feature views) via majority voting or weighted fusion. **String kernels** (character-spectrum kernels) were particularly strong baselines for NLI, consistent with results in authorship attribution tasks.

Results

Closed task (TOEFL11 only)

Top accuracy: 0.836 (JAR). Next: OSL 0.834; BUC 0.827; CAR 0.826; TUE 0.822. The closed task had **29 teams** and **116 submissions**.

Open-1 (external data only)

Performance dropped with domain mismatch. **Top accuracy:** TOR 0.565; followed by TUE 0.385; NAI 0.356 — illustrating that *genre and data match matter more than sheer data volume* when transferring to TOEFL-style essays.

Open-2 (TOEFL11 + external)

Adding external data helped when combined with in-domain TOEFL11. **Top accuracy:** TUE 0.835; TOR 0.816; HYD 0.741; NAI 0.703.

Post-task 10-fold cross-validation (TRAIN+DEV)

On unified TRAIN+DEV (10-fold CV), best accuracies clustered around mid-80s: **CN 84.6**, **JAR 84.5**, **OSL 83.9**, **BUC 82.6**, **MQ 82.5**, **TUE 82.4**, **CAR 82.2**, **NAI 82.1**. For context, Tetreault et al. (2012) reported **80.9** on a comparable setup.

Quick view — Accuracies

Subtask	Top Team	Accuracy	Next Best
Closed	JAR	0.836	OSL 0.834; BUC 0.827; CAR 0.826
Open-1	TOR	0.565	TUE 0.385; NAI 0.356
Open-2	TUE	0.835	TOR 0.816; HYD 0.741; NAI 0.703

Subtask	Top Team	Accuracy	Next Best
10-fold CV (TRAIN+DEV)	CN	0.846	JAR 0.845; OSL 0.839; BUC 0.826

Key takeaways

1) **Surface features win**: character/word/POS n-grams carry most of the signal. 2) **Data match > data size**: external data without genre alignment hurts; adding it to TOEFL11 helps modestly. 3) **SVMs remain hard to beat** for sparse, high-dimensional text features. 4) Benchmarks and public splits enable real progress and honest comparisons.

Ideas for future work

Re-run NLI with modern encoders (e.g., XLM-R, DeBERTa-v3) and character-aware CNN/RNN submodules; compare to strong SVM+string-kernel baselines. Expand beyond English L2 to Japanese L2 (useful in Japan-focused EdTech). Integrate explicit error annotations to analyze which error classes contribute most to L1 discrimination.

Reference

Tetreault, J., Blanchard, D., & Cahill, A. (2013). *A Report on the First Native Language Identification Shared Task*. In *BEA@NAACL-HLT* (pp. 48–57). ACL Anthology.