# Chinese Native Language Identification

**Shervin Malmasi**
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
shervin.malmasi@mq.edu.au

**Mark Draس**

## Abstract

We present the first application of Native Language Identification (NLI) to non-English data. Motivated by theories of language transfer, NLI is the task of identifying a writer's native language (L1) based on their writings in a second language (the L2). An NLI system was applied to Chinese learner texts using topic-independent syntactic models to assess their accuracy. We find that models using part-of-speech tags, context-free grammar production rules and function words are highly effective, achieving a maximum accuracy of 71% . Interestingly, we also find that when applied to equivalent English data, the model performance is almost identical. This finding suggests a systematic pattern of cross-linguistic transfer may exist, where the degree of transfer is independent of the L1 and L2.

## 1 Introduction

Native Language Identification (NLI) is the task of identifying an author's native language (L1) based on their writings in a second language (the L2). NLI works by identifying language use patterns that are common to groups of speakers that share the same native language. This process is underpinned by the presupposition that an author's L1 will dispose them towards particular language production patterns in their L2, as influenced by their mother tongue. This relates to Cross-Linguistic Influence (CLI), a key topic in the field of Second Language Acquisition (SLA) that analyzes transfer effects from the L1 on later learned languages (Ortega, 2009).

While NLI has applications in security, most research has a strong linguistic motivation relating to language teaching and learning. Rising numbers of language learners have led to an increasing need for language learning resources, which has in turn fuelled much of the language acquisition research of the past decade. In this context, by identifying L1-specific language usage and error patterns, NLI can be used to better understand SLA and develop teaching methods, instructions and learner feedback that is specific to their mother tongue.

However, all of the NLI research to date has focused exclusively on English L2 data. To this end there is a need to apply NLI to other languages, not only to gauge their applicability but also to aid in teaching research for other emerging languages.

Interest in learning Chinese is rapidly growing, leading to increased research in Teaching Chinese as a Second Language (TCSL) and the development of related resources such as learner corpora (Chen et al., 2010). The application of these tools and scientific methods like NLI can greatly assist researchers in creating effective teaching practices and is an area of active research.

The aim of this research is to evaluate the cross-language applicability of NLI techniques by applying them to Chinese learner texts, evaluating their efficacy and comparing the results with their English equivalents.

To the best of our knowledge this is the first reported application of NLI to non-English data and we believe this is an important step in gaining deeper insights about the technique.

## 2 Related Work

NLI is a fairly recent, but rapidly growing area of research. While some research was conducted in the early 2000s, the most significant work has only appeared in the last few years (Wong and Dras, 2009; Wong and Dras, 2011; Swanson and Charniak, 2012; Tetreault et al., 2012; Bykh and Meurers, 2012).

Most studies approach NLI as a multi-class supervised classification task. In this experimental design, the L1 metadata are used as class labels

*Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 95–99,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

and the individual writings are used as training and testing data. Using lexical and syntactic features of increasing sophistication, researchers have obtained good results under this paradigm. While a detailed exposition of NLI has been omitted here due to space constraints, a concise review can be found in Bykh and Meurers (2012).

### 2.1 NLI 2013 Shared Task

This increased interest brought unprecedented level of research focus and momentum, resulting in the first NLI shared task being held in 2013.[1] The shared task aimed to facilitate the comparison of results by providing a large NLI-specific dataset and evaluation procedure, to enable direct comparison of results achieved through different methods. Overall, the event was considered a success, drawing 29 entrants and experts from not only Computational Linguistics, but also SLA. The best teams achieved accuracies of greater than 80% on this 11-class classification task. A detailed summary of the results is presented in Tetreault et al. (2013).

## 3 Data

Growing interest has led to the recent development of the Chinese Learner Corpus (Wang et al., 2012), the first large-scale corpus of learner texts comprised of essays written by university students. Learners from 59 countries are represented and proficiency levels have been sampled representatively across beginners, intermediate and advanced learners. However, texts by native speakers of other Asian countries are disproportionately represented, likely due to geographical proximity.

For this work we extracted 3.75 million tokens of text from the CLC in the form of individual sentences.[2] Following the methodology of Brooke and Hirst (2011), we combine the sentences from the same L1 to form texts of 600 tokens on average, creating a set of documents suitable for NLI[3].

We choose the top 11 languages, shown in Table 1, to use in our experiments. This is due to two considerations. First, while many L1s are represented in the corpus, most have relatively few texts. Choosing the top 11 classes allows us to balance having a large number of classes, and also maximizes the amount of data used. Secondly, this is the same number of classes used in the NLI 2013 shared task, enabling us to draw cross-language comparisons with the shared task results.

| Language | Size | Language | Size |
|---|---|---|---|
| Filipino FIL | 415 | Indonesian IND | 402 |
| Thai THA | 400 | Laotian LAO | 366 |
| Burmese MYA | 349 | Korean* KOR | 330 |
| Khmer KHM | 294 | Vietnamese VIE | 267 |
| Japanese* JAP | 180 | Spanish* SPA | 112 |
| Mongolian MON | 101 | | |

Table 1: Our data, broken down by language and the number of texts in each class. Languages overlapping with the TOEFL11 corpus marked with *.

## 4 Experimental Setup

We also follow the supervised classification approach described in §2. We devise and run experiments using several models that capture different types of linguistic information. For each model, features are extracted from the texts and a classifier is trained to predict the L1 labels using the features. As our data is not topic-balanced, we avoid using topic-dependent lexical features such as character or word $n$-grams.

Each experiment is run with two feature representations: binary (presence/absence of a feature) and normalized frequencies, where feature values are normalized to text length using the $l2$-norm.

### 4.1 Parser

The Stanford CoreNLP[4] suite of NLP tools and the provided Chinese models are used to tokenize, PoS tag and parse the unsegmented corpus texts.

### 4.2 Classifier

We use Support Vector Machines for classification. Specifically, we use the LIBLINEAR SVM package (Fan et al., 2008) as it is well-suited to text classification tasks with large numbers of features and texts. We use the L2-regularized L2-loss support vector classification (dual) solver.

### 4.3 Evaluation

The same evaluation metrics and standards used in the NLI2013 Shared Task are used: we report classification accuracy under 10-fold cross-validation. We also use the same number of classes as the shared task to facilitate comparative analyses.

[1] Organised by the Educational Testing Service and co-located with the eighth instalment of the Building Educational Applications Workshop at NAACL/HLT 2013. sites.google.com/site/nlisharedtask2013/

[2] Full texts are not made available, only individual sentences with the relevant metadata (proficiency/nationality).

[3] Pending permission from the CLC corpus authors, we will attempt to release the Chinese NLI dataset publicly.

[4] http://nlp.stanford.edu/software/corenlp.shtml

| **Feature** | **Accuracy (%)** | |
|---|---|---|
| | Binary | Frequency |
| Random Baseline | 9.09 | 9.09 |
| PoS unigrams | 20.12 | 35.32 |
| Part-of-Speech bigrams | 32.83 | 54.24 |
| Part-of-Speech trigrams | 47.24 | 55.60 |
| Function Words | 43.93 | 51.91 |
| Production Rules | 36.14 | 49.80 |
| All features | 61.75 | 70.61 |

Table 2: Chinese Native Language Identification accuracy (%) for all of our models.

## 5 Experiments and Results

### 5.1 Part-of-Speech tag $n$-grams

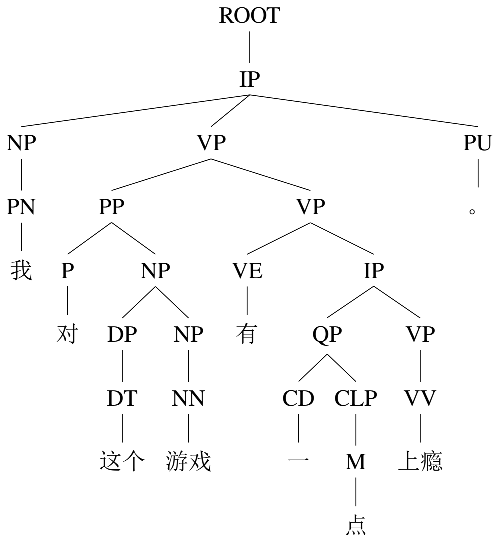
Our first experiment assesses the utility of the syntactic information captured by part-of-speech (PoS) tags for Chinese NLI. The PoS tags for each text are predicted and $n$-grams of size 1–3 are extracted from the tags. These $n$-grams capture (very local) syntactic patterns of language use and are used as classification features.

The results for these three features, and our other models are shown in Table 2. The trigram frequencies give the best accuracy of 55.60%, suggesting that there exist group-specific patterns of Chinese word order and category choice which provide a highly discriminative cue about the L1.

### 5.2 Function Words

As opposed to content words, function words are topic-independent grammatical words that indicate the relations between other words. They include determiners, conjunctions and auxiliary verbs. Distributions of English function words have been found to be useful in studies of authorship attribution and NLI. Unlike PoS tags, this model analyzes the author's specific word choices.

We compiled a list of 449 Chinese function words[5] to be used as features in this model. As shown in Table 2, the function word frequency features provide the best accuracy of 51.91%, significantly higher than the random baseline. This again suggests the presence of L1-specific grammatical and lexical choice patterns that can help distinguish the L1, potentially due to cross-linguistic transfer. Such lexical transfer effects have been previously noted by researchers and linguists (Odlin, 1989). These effects are mediated not only by cognates and similarities in word forms, but also word semantics and meanings.



```
IP → NP VP PU        VP → PP VP
NP → DP NP           PP → P NP
```

Figure 1: A constituent parse tree for a sentence from the corpus along with some of the context-free grammar production rules extracted from it.

### 5.3 Context-free Grammar Production Rules

In the next experiment we investigate the differences in the distribution of the context-free grammar production rules used by the learners. To do this, constituent parses for all sentences are obtained and the production rules, excluding lexicalizations, are extracted. Figure 1 shows a sample tree and rules. These context-free phrase structure rules capture the overall structure of grammatical constructions and are used as classification features in this experiment.

As seen in Table 2, the model achieves an accuracy of 49.80%. This supports the hypothesis that the syntactic substructures contain characteristic constructions specific to L1 groups and that these syntactic cues strongly signal the writer's L1.

### 5.4 Combining All Features

Finally, we assess the redundancy of the information captured by our models by combining them all into one vector space to create a single classifier. From Table 2 we see that for each feature representation, the combined feature results are higher than the single best feature, with a max-

[5] The function word list was compiled from Chinese language teaching resources. The complete list can be accessed at http://comp.mq.edu.au/~madras/research/data/chinese-fw.txt

imum accuracy of 70.61%. This demonstrates that for at least some of the features, the information they capture is orthogonal and complementary, and combining them can improve results.

## 6 Discussion

A key finding here is that NLI models can be successfully applied to non-English data. This is an important step for furthering NLI research as the field is still relatively young and many fundamental questions have yet to be answered.

All of the tested models are effective, and they appear to be complementary as combining them improves overall accuracy. We also note the difference in the efficacy of the feature representations and see a clear preference for frequency-based feature values. Others have found that binary features are the most effective for English NLI (Brooke and Hirst, 2012), but our results indicate frequency information is more informative in this task. The combination of both feature types has also been reported to be effective (Malmasi et al., 2013).

To see how these models perform across languages, we also compare the results against the TOEFL11 corpus used in the NLI2013 shared task. We perform the same experiments on that dataset using the English CoreNLP models, Penn Treebank PoS tagset and a set of 400 English function words. Figure 2 shows the results side by side.

Remarkably, we see that the model results closely mirror each other across corpora. This is a highly interesting finding from our study that merits further investigation. There is a systematic pattern occurring across data from learners of completely different L1-L2 pairs. This suggests that manifestations of CLI via surface phenomena occur at the same levels and patternings regardless of the L2. Cross-language studies can help researchers in linguistics and cognitive science to better understand the SLA process and language transfer effects. They can enhance our understanding of how language is processed in the brain in ways that are not possible by just studying monolinguals or single L1-L2 pairs, thereby providing us with important insights that increase our knowledge and understanding of the human language faculty.

One limitation of this work is the lack of similar amounts of training data for each language. However, many of the early and influential NLI studies (e.g. Koppel et al. (2005), Tsur and Rappoport (2007)) were performed under similar circumstances. This issue was noted at the time, but did not deter researchers as corpora with similar issues were used for many years. Non-English NLI is also at a similar state where the extant corpora are not optimal for the task, but no other alternatives exist for conducting this research.

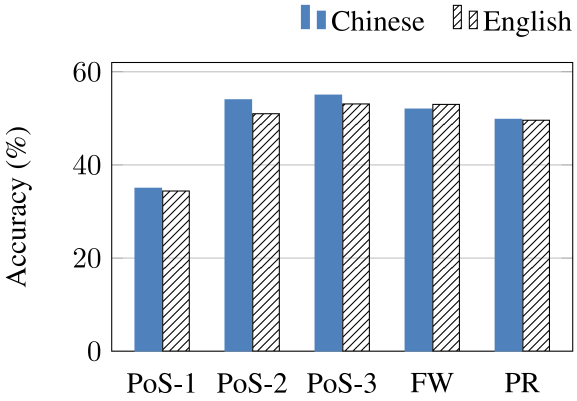


Figure 2: Comparing feature performance on the Chinese Learner Corpus and English TOEFL11 corpora. PoS-1/2/3: PoS uni/bi/trigrams, FW: Function Words, PR: Production Rules

Finally, there are also a number of way to further develop this work. Firstly, the experimental scope could be expanded to use even more linguistically sophisticated features such as dependency parses. Model accuracy could potentially be improved by using the metadata to develop proficiency-segregated models. Classifier ensembles could also help in increasing the accuracy.

## 7 Conclusion

In this work we have presented the first application of NLI to non-English data. Using the Chinese Learner Corpus, we compare models based on PoS tags, function words and context-free grammar production rules and find that they all yield high classification accuracies.

Comparing the models against an equivalent English learner corpus we find that the accuracies are almost identical across both L2s, suggesting a systematic pattern of cross-linguistic transfer where the degree of transfer is independent of the L1 and L2. Further research with other L2 learner corpora is needed to investigate this phenomena.

## Acknowledgments

We wish to thank Associate Professor Maolin Wang for providing access to the CLC corpus, and Zhendong Zhao for his assistance. We also thank the reviewers for their constructive feedback.

## References

Julian Brooke and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. In *Conference of Learner Corpus Research (LCR2011)*, Louvain-la-Neuve, Belgium. Presses universitaires de Louvain.

Julian Brooke and Graeme Hirst. 2012. Robust, Lexicalized Native Language Identification. In *Proceedings of COLING 2012*, pages 391–408, Mumbai, India, December. The COLING 2012 Organizing Committee.

Serhiy Bykh and Detmar Meurers. 2012. Native Language Identification using Recurring $n$-grams – Investigating Abstraction and Domain Dependence. In *Proceedings of COLING 2012*, pages 425–440, Mumbai, India, December. The COLING 2012 Organizing Committee.

Jianguo Chen, Chuang Wang, and Jinfa Cai. 2010. *Teaching and learning Chinese: Issues and perspectives*. IAP.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*, volume 3495 of *LNCS*, pages 209–217. Springer-Verlag.

Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. 2013. Nli shared task 2013: Mq submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June. Association for Computational Linguistics.

Terence Odlin. 1989. *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge University Press, Cambridge, UK.

Lourdes Ortega. 2009. *Understanding Second Language Acquisition*. Hodder Education, Oxford, UK.

Benjamin Swanson and Eugene Charniak. 2012. Native Language Detection with Tree Substitution Grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 193–197, Jeju Island, Korea, July. 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, December. The COLING 2012 Organizing Committee.

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, June. Association for Computational Linguistics.

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

Maolin Wang, Qi Gong, Jie Kuang, and Ziyu Xiong. 2012. The development of a chinese learner corpus. In *Speech Database and Assessments (Oriental COCOSDA), 2012 International Conference on*, pages 1–6. IEEE.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive Analysis and Native Language Identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia, December.

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., July. Association for Computational Linguistics.