Finnish Native Language Identification

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

Mark Dras
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
mark.dras@mq.edu.au

Abstract

We outline the first application of Native Language Identification (NLI) to Finnish learner data. NLI is the task of predicting an author's first language using writings in an acquired language. Using data from a new learner corpus of Finnish — a language typology quite different from others previously investigated, with its morphological richness potentially causing difficulties — we show that a combination of three feature types is useful for this task. Our system achieves an accuracy of 70% against a baseline of 20% for predicting an author's L1. Using the same features we can also distinguish non-native writings with an accuracy of 97%. This methodology can be useful for studying language transfer effects, developing teaching materials tailored to students' native language and also forensic linguistics.

1 Introduction

It has been noted in the linguistics literature since the 1950s that speakers of particular languages have characteristic production patterns when writing in a second language. This language transfer phenomenon has been investigated independently in a number of fields from different perspectives, including qualitative research in Second Language Acquisition (SLA) and more recently though predictive computational models in NLP (Jarvis and Crossley, 2012).

Such analyses have traditionally been conducted manually by researchers, and the issues that arise when they are attempted on large corpora are well known (Ellis, 2008). Recently, researchers have noted that NLP has the tools to use large amounts of data to automate this analysis, using complex feature types. This has motivated studies in Native Language Identification (NLI), a subtype of text classification where the goal is to determine the native language (L1) of an author using texts they have written in a second language or L2 (Tetreault et al., 2013).

Most work in SLA, NLI and NLP for that matter has dealt with English. This is largely due to the fact that since World War II, the world has witnessed the ascendancy of English as its *lingua franca*. While English is the native language of over 400 million people in the U.S., U.K. and the Commonwealth, there are also over a billion people who speak English as their second or foreign language (Guo and Beckett, 2007). This has created a global environment where learning multiple languages is not exceptional and this has fueled the growing research into language acquisition.

However, while English is one of the most prevalent languages in the world there are still a sizeable number of jobs and activities in parts of the world where the acquisition of a language other than English is a necessity.

One such example is Finland, where due to the predicted labour shortage, the government has adopted policies encouraging economic and work-related migration (Ministry of Labour, 2006), with an emphasis on the role of the education system. Aiding new immigrants to learn the Finnish language has been a key pillar of this policy particularly as learning the language of the host nation has been found to be an important factor for social integration and assimilation (Nieminen, 2009). This, in turn, has motivated research in studying the acquisition of Finnish to identify the most challenging aspects of the process.[1]

Finnish differs from English in many respects including verb tenses and forms (Karlsson, 2008). It is a highly inflectional agglutinative language with a flexible word order.[2]

[1] For example, the recent study by Siitonen (2014)

[2] More information about these differences can be found at http://esl.fis.edu/grammar/langdiff/

Shervin Malmasi and Mark Dras. 2014. Finnish Native Language Identification. In *Proceedings of Australasian Language Technology Association Workshop*, pages 139−144.

Given these differences, the main objective of the present study is to determine if NLI techniques previously applied to L2 English can be effective for detecting L1 transfer effects in L2 Finnish.

2 Background

NLI is a fairly recent, but rapidly growing area of research. While some early research was conducted in the early 2000s, most work has only appeared in the last few years. This surge of interest, coupled with the inaugural shared task in 2013 have resulted in NLI becoming a well-established NLP task. The NLI Shared Task in 2013 was attended by 29 teams from the NLP and SLA areas. An overview of the shared task results and a review of prior NLI work can be found in Tetreault et al. (2013).

While there exists a large body of literature produced in the last decade, almost all of this work has focused exclusively on L2 English. The most recent work in this field has successfully presented the first applications of NLI to a large non-English datasets (Malmasi and Dras, 2014b; Malmasi and Dras, 2014a), evidencing the usefulness of syntactic features in distinguishing L2 Chinese and L2 Arabic texts.

Finnish poses a particular challenge. In terms of morphological complexity, it is among the world's most extreme: its number of cases, for example, places it in the highest category in the comparative World Atlas of Language Structures (Iggesen, 2013). Comrie (1989) proposed two scales for characterising morphology, the index of synthesis (based on the number of categories expressed per morpheme) and the index of fusion (based on the number of categories expressed per morpheme). While an isolating language like Vietnamese would have an index of synthesis score close to 1, the lowest possible score, Finnish scores particularly high on this metric (Pirkola, 2001). Because of this morphological richness, and because it is typically associated with freeness of word order, Finnish potentially poses a problem for the quite strongly lexical features currently used in NLI.

3 Data

Although the majority of currently available learner corpora are based on English L2 (Granger, 2012), data collection from learners of other languages such as Finnish has also attracted attention in recent years.

Native Language	Documents
Russian	40
Japanese	34
Lithuanian	28
Czech	27
German	21
Hungarian	21
Polish	12
Komi	11
English	10
Total	204

Table 1: The L1 classes included in this experiment and the number of texts within each class.

The present study is based on texts from the Corpus of Advanced Learner Finnish (LAS2) which is comprised of L2 Finnish writings (Ivaska, 2014). The texts are being collected as part of an ongoing project at the University of Turku[3] since 2007 with the goal of collection suitable data than allows for quantitative and qualitative analysis of Finnish interlanguage.

The current version of the corpus contains approximately 630k tokens of text in 640 texts collected from writers of 15 different L1 backgrounds. The included native language backgrounds are: Czech, English, Erzya, Estonian, German, Hungarian, Icelandic, Japanese, Komi, Lithuanian, Polish, Russian, Slovak, Swedish and Udmurt. The corpus texts are available in an XML format and have been annotated in terms of parts of speech, word lemmas, morphological forms and syntactic functions.

While there are 15 different L1s represented in the corpus, the majority of these have less than 10 texts and cannot reliably be used for NLI. Instead we use a subset of the corpus consisting of the top seven native languages by number of texts. The languages and document counts in each class are shown in Table 1.

4 Experimental Methodology

In this study we employ a supervised multi-class classification approach. The learner texts from

finnish.htm

[3] http://www.utu.fi/fi/yksikot/hum/yksikot/suomi-sgr/tutkimus/tutkimushankkeet/las2/Sivut/home.aspx

the corpus are organized into classes according on the author's L1 and these documents are used for training and testing in our experiments.

4.1 Classifier

We use a linear Support Vector Machine to perform multi-class classification in our experiments. In particular, we use the LIBLINEAR[4] package (Fan et al., 2008) which has been shown to be efficient for text classification problems such as this. More specifically, it has been demonstrated to be the most effective classifier for this task in the 2013 NLI Shared Task (Tetreault et al., 2013).

4.2 Evaluation Methodology

Consistent with most previous NLI studies and the NLI 2013 shared task, we report results as classification accuracy under k-fold cross-validation, with $k = 10$. In recent years this has become a *de facto* standard for reporting NLI results.

5 Experiments

We experiment using three different feature types described in this section. Previous NLI research on English data has utilized a range of features types varying from surface features to more sophisticated syntactic ones (Malmasi et al., 2013). However, in most such studies the use of such deeper features is predicated on the availability of NLP tools and models for extracting those features. This, unfortunately, is not the case for Finnish and it was decided to make use of a simpler feature set in this preliminary study.

As our data is not balanced for topic, we do not consider the use of purely lexical features such as word n-grams in this study. Topic bias can occur as a result of the subject matters or topics of the texts to be classified not not evenly distributed across the classes. For example, if in our training data all the texts written by English L1 speakers are on topic A, while all the French L1 authors write about topic B, then we have implicitly trained our classifier on the topics as well. In this case the classifier learns to distinguish our target variable through another confounding variable. Others researchers like Brooke and Hirst (2012), however, argue that lexical features cannot be simply ignored. Given the small size of our data and the inability to reach definitive conclusions regarding this, we do not attempt to explore this issue here.

5.1 Finnish Function Words

The distributions of grammatical function words such as determiners and auxiliary verbs have proven to be useful in NLI. This is considered to be a useful syntactic feature as these words indicate the relations between content words and are topic independent. The frequency distributions of 700 Finnish function words[5] were extracted from the learner texts and used as features in this model.

5.2 Part-of-Speech tag n-grams

In this model POS n-grams of size 1–3 were extracted. These n-grams capture small and very local syntactic patterns of language production and were used as classification features. Previous work and our experiments showed that sequences of size 4 or greater achieve lower accuracy, possibly due to data sparsity, so we do not include them.

5.3 Character n-grams

This is a sub-lexical feature that uses the constituent characters that make up the whole text. From a linguistic point of view, the substrings captured by this feature, depending on the order, can implicitly capture various sub-lexical features including letters, phonemes, syllables, morphemes and suffixes. We do not consider n-grams of order 4 or higher as they may be capturing whole words.

5.4 Identifying Non-Native Writing

Our final experiment involves using the above-described features to classify Finnish texts as either Native or non-Native. To achieve this we use 100 control texts included in the LAS2 corpus that written by native Finnish speakers to represent the Native class. This is contrasted against the non-Native class which includes 100 texts sampled from each language[6] listed in Table 1.

6 Results

The results of the first three experiments are shown in Table 2. The majority baseline is calculated by using the largest class, in this case Russian,[7] as the

[4] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[5] These were sourced from pre-existing word lists from http://members.unine.ch/jacques.savoy/clef/index.html

[6] English only has 10 texts, so we include 2 extra Japanese texts to create a set of 100 documents.

[7] $40/204 = 19.6\%$

Feature	Accuracy (%)
Majority Baseline	19.6
Character unigrams	34.8
Character bigrams	42.6
Character trigrams	53.9
Function Words	54.6
Part-of-Speech unigrams	36.3
Part-of-Speech bigrams	55.2
Part-of-Speech trigrams	54.8
All features combined	**69.5**

Table 2: Finnish Native Language Identification accuracy for the three experiments in this study.

Feature	Accuracy (%)
Chance Baseline	50.0
Character unigrams	91.0
Character bigrams	94.0
Character trigrams	95.0
Function Words	94.0
Part-of-Speech unigrams	88.0
Part-of-Speech bigrams	89.5
Part-of-Speech trigrams	91.5
All features combined	**97.0**

Table 3: Accuracy for classifying texts as Native or non-Native (Experiment 4).

default classification label chosen for all texts. No other baselines are available here since this is the first NLI work using this data and L2 language.

The character n-gram models all perform well-above the baseline, with higher accuracies as n increases. Similarly, the distribution of function words is highly discriminative, yielding 54.6% accuracy. The purely syntactic POS n-gram models are also very useful for this task, with the best accuracy of 54.8% for POS trigrams.

Combining all of the models into a single feature vector provides the highest accuracy of 69.5%, around 15% better than the best single feature type. This demonstrates that the information captured by the various models is complementary and that the feature types are not redundant.

The results of our final experiment for distinguishing non-Native writing are listed in Table 3. They demonstrate that these feature types are highly useful for discriminating between Native and non-Native writings, achieving 97% accuracy by using all feature types. Character trigrams are the best single feature in this experiment.

7 Discussion

The most significant finding here is that the NLI methodology can be successfully applied to Finnish data with results that are largely comparable to state-of-the-art English NLI systems.

The main contributions of this work include the identification of a new dataset for NLI and employing it to demonstrate the cross-linguistic nature of NLI. This is one of the very first applications of NLI to a language other than English and an important step in the growing field of NLI, particularly with the current drive to investigate other languages.

NLI technology has practical applications in various fields. One potential application is in the field of forensic linguistics (Coulthard and Johnson, 2007), a juncture where the legal system and linguistic stylistics intersect (Gibbons and Prakasam, 2004). Here NLI can be used as a tool for Authorship Profiling (Grant, 2007) to provide evidence about a writer's linguistic background. There are a number of situations where a text, like an anonymous letter, is the key piece of evidence in an investigation. Clues about the native language of a writer can help investigators in identifying the source.[8] Accordingly, we can see that NLI can be a useful forensic tool for law enforcement agencies. In fact, recent NLI research such as that related to the work presented by (Perkins, 2014) has already attracted interest and funding from intelligence agencies (Perkins, 2014, p. 17).

In addition to applications in forensic linguistics, NLI can aid the development of research tools for SLA researchers investigating language transfer and cross-linguistic effects. Similar data-driven methods have been recently applied to generate potential language transfer hypotheses from the writings of English learners (Swanson and Charniak, 2014; Malmasi and Dras, 2014d). By using an error annotated corpus, which was not the case in this study, the annotations could be used in conjunction with similar linguistic features to study the syntactic contexts in which different error types occur (Malmasi and Dras, 2014c). Results from such approaches could be used to create teaching material that is customized for the

[8] e.g. for analysing extremist related activity on the web (Abbasi and Chen, 2005)

learner's L1. This has been previously shown to yield learning improvements (Laufer and Girsai, 2008).

There are a number of avenues for future work. A key limitation of this study, although beyond our control, is the limited amount of data used. We hope to evaluate our system on larger data as it becomes available. The application of more linguistically sophisticated features also merits further investigation, but this is limited by the availability of Finnish NLP tools and resources. Another possible improvement is the use of classifier ensembles to improve classification accuracy. This has previously been applied to English NLI with good results (Tetreault et al., 2012).

We would also like to point to the failure to distinguish between the L2 and any other acquired languages as a more general criticism of the NLI literature to date. The current body of NLI literature fails to distinguish whether the learner language is in fact the writer's second language, or whether it is possibly a third language (L3).

It has been noted in the SLA literature that when acquiring an L3, there may be instances of both L1- and L2-based transfer effects on L3 production (Ringbom, 2001). Studies of such second language transfer effects during third language acquisition have been a recent focus on cross-linguistic influence research (Murphy, 2005).

One potential reason for this shortcoming in NLI is that none of commonly used corpora distinguish between the L2 and L3; they only include the author's L1 and the language which they are learning. This language is generally assumed to be an L2, but this may not be case. At its core, this issue relates to corpus linguistics and the methodology used to create learner corpora. The thorough study of these effects is contingent upon the availability of more detailed language profiles of authors in learner corpora. The manifestation of these interlanguage transfer effects (the influence of one non-native language on another) are dependent on the status, recency and proficiency of the learner's acquired languages (Cenoz and Jessner, 2001). Accordingly, these variables need to be accounted for by the corpus creation methodology.

But it should also be noted that based on currently available evidence, identifying the specific source of cross-linguistic influence in speakers of an L3 or additional languages (L4, L5, etc.) is not an easy task. Recent studies point to the methodological problems in studying productions of multilinguals (De Angelis, 2005; Williams and Hammarberg, 1998; Dewaele, 1998).

From an NLP standpoint, if the author's acquired languages or their number is known, it may be possible to attempt to trace different transfer effects to their source using advanced segmentation techniques. We believe that this is an interesting task in itself and a potentially promising area of future research.

References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group Web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.

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.

Jasone Cenoz and Ulrike Jessner. 2001. *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*, volume 31. Multilingual Matters.

Bernard Comrie. 1989. *Language Universals and Linguistic Typology*. University of Chicago Press, Chicago, IL, US, 2nd edition.

Malcolm Coulthard and Alison Johnson. 2007. *An introduction to Forensic Linguistics: Language in evidence*. Routledge.

Gessica De Angelis. 2005. Multilingualism and non-native lexical transfer: An identification problem. *International Journal of Multilingualism*, 2(1):1–25.

Jean-Marc Dewaele. 1998. Lexical inventions: French interlanguage as L2 versus L3. *Applied Linguistics*, 19(4):471–490.

Rod Ellis. 2008. *The Study of Second Language Acquisition, 2nd edition*. Oxford University Press, Oxford, UK.

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.

John Gibbons and Venn Prakasam. 2004. *Language in the Law*. Orient Blackswan.

Sylviane Granger. 2012. Learner corpora. *The Encyclopedia of Applied Linguistics*.

Tim Grant. 2007. Quantifying evidence in forensic authorship analysis. *International Journal of Speech Language and the Law*, 14(1):1–25.

Yan Guo and Gulbahar H Beckett. 2007. The hegemony of english as a global language: Reclaiming local knowledge and culture in china. *Convergence*, 40:117–132.

Oliver A. Iggesen, 2013. *Number of Cases*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ilmari Ivaska. 2014. The corpus of advanced learner Finnish (LAS2): database and toolkit to study academic learner Finnish. *Apples*, 8.

Scott Jarvis and Scott Crossley, editors. 2012. *Approaching Language Transfer Through Text Classification: Explorations in the Detection-based Approach*, volume 64. Multilingual Matters Limited, Bristol, UK.

Fred Karlsson. 2008. *Finnish: An essential grammar*. Routledge.

Batia Laufer and Nany Girsai. 2008. Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, 29(4):694–716.

Shervin Malmasi and Mark Dras. 2014a. Arabic Native Language Identification. In *Proceedings of the Arabic Natural Language Processing Workshop (colocated with EMNLP 2014)*, pages 180–186, Doha, Qatar, October. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2014b. Chinese Native Language Identification. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, April.

Shervin Malmasi and Mark Dras. 2014c. From Visualisation to Hypothesis Construction for Second Language Acquisition. In *Graph-Based Methods for Natural Language Processing*, pages 56–64, Doha, Qatar, October. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2014d. Language Transfer Hypotheses with Linear SVM Weights. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1385–1390.

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.

Ministry of Labour. 2006. Hallituksen maahanmuuttopoliittinen ohjelma. *Tyhallinnon julkaisu 371*.

Shirin Murphy. 2005. Second language transfer during third language acquisition. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 3(1).

Tanja Nieminen. 2009. Becoming a new Finn through language: non-native English-speaking immigrants' views on integrating into Finnish society.

Ria Perkins. 2014. *Linguistic identifiers of L1 Persian speakers writing in English: NLID for authorship analysis*. Ph.D. thesis, Aston University.

Ari Pirkola. 2001. Morphological typology of languages for IR. *Journal of Documentation*, 57(3):330–348.

Hakan Ringbom. 2001. Lexical transfer in L3 production. (Cenoz and Jessner, 2001), pages 59–68.

Kirsti Siitonen. 2014. Learners' dilemma: an example of complexity in academic Finnish. The frequency and use of the E infinitive passive in L2 and L1 Finnish. *AFinLA-e: Soveltavan kielitieteen tutkimuksia*, (6):134–148.

Ben Swanson and Eugene Charniak. 2014. Data Driven Language Transfer Hypotheses. *EACL 2014*, page 169.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, Beata Beigman-Klebanov, and Martin Chodorow. 2012. Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification. In *Proc. Internat. Conf. on Computat. Linguistics (COLING)*.

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.

Sarah Williams and Bjorn Hammarberg. 1998. Language switches in L3 production: Implications for a polyglot speaking model. *Applied linguistics*, 19(3):295–333.