Using Classifier Features to Determine Language Transfer on Morphemes

Alexandra Lavrentovich
Department of Linguistics, University of Florida
Gainesville, FL 32601 USA
alavrent@ufl.edu

Abstract

The aim of this thesis is to perform a Native Language Identification (NLI) task where we identify an English learner's native language background based only on the learner's English writing samples. We focus on the use of English grammatical morphemes across four proficiency levels. The outcome of the computational task is connected to a position in second language acquisition research that holds all learners acquire English grammatical morphemes in the same order, regardless of native language background. We use the NLI task as a tool to uncover cross-linguistic influence on the developmental trajectory of morphemes. We perform a cross-corpus evaluation across proficiency levels to increase the reliability and validity of the linguistic features that predict the native language background. We include native English data to determine the different morpheme patterns used by native versus non-native English speakers. Furthermore, we conduct a human NLI task to determine the type and magnitude of language transfer cues used by human judges versus the classifier.

1 Introduction

Native Language Identification (NLI) is a text classification task that determines the native language background (L1) of a writer based solely on the writer's second language (L2) production. Within computational linguistics, there is a flurry of activity as researchers test the best classifier models with a wide range of linguistic and stylistic features to reach the highest classification accuracy scores. NLI is pursued in a variety of written genres such as argumentative essays (Crossley and McNamara, 2012), online journal entries (Brooke and Hirst, 2012), scientific articles (Stehwien and Padó, 2015) as well as transcribed spoken language (Zampieri et al., 2017), and even eye fixation data while reading (Berzak et al., 2017). Although the majority of NLI studies have used English as the L2, recent work expands the NLI task to other L2 languages including Arabic, Chinese, and Spanish (Malmasi, 2016).

These forays illustrate that computational methods are robust for identifying the first language background of the language learner. This task has interesting implications for exploring how the L1 permeates into the L2. Specifically, by selecting and analyzing the most informative linguistic features that predict the L1 background, we can generate testable hypotheses about language transfer that would be of value to second language acquisition (SLA) researchers and practitioners. One such application would be to determine if grammatical morphemes are susceptible to cross-linguistic influence (CLI) throughout development, which have previously been described to be learned in the same order, regardless of the L1.

The thesis consists of four main studies. First, we conduct the NLI task and feature analysis for ten L1 backgrounds across four proficiency levels to determine the nature and extent of CLI on the developmental trajectory of morpheme production. Second, we include native English data in a second iteration of the NLI task to determine the linguistic patterns that vary between native and non-native English writers. Third, we conduct a cross-corpus evaluation across proficiency levels to determine which features are more reliable and corpus-independent. Fourth, we conduct a human NLI task to determine the linguistic cues used by humans versus machines in detecting a writer's L1. Taking these studies together, the thesis uses NLI to support and inform topics in SLA, and conversely, we connect principles in SLA to NLI by expanding and building new NLI models.

Proceedings of NAACL-HLT 2018: Student Research Workshop, pages 61–66
New Orleans, Louisiana, June 2 - 4, 2018.

Stage	Morpheme
1	Progressive *-ing*
	Plural *-s*
	Copula *be*
2	Auxiliary *be*
	Articles *a/an/the*
3	Irregular past
4	Regular past *-ed*
	Third person singular *-s*
	Possessive *'s*

Table 1: Natural order of English L2 morpheme acquisition.

2 Background

We take the NLI task as a starting point for investigating CLI on the developmental trajectory of English grammatical morphemes. In particular, we determine the patterns of overuse, underuse, and erroneous use of morphemes by learners from ten L1 backgrounds across four proficiency levels. We focus on morphemes in particular in order to connect to SLA research that suggests English learners acquire English grammatical morphemes in a universal order, regardless of the learner's L1 background (Dulay and Burt, 1974; Larsen-Freeman, 1975). Krashen (1985) advocated for a shared order composed of four acquisition stages, illustrated in Table 1.

More recently, studies have identified multiple determinants predicting the shared order of accurate morpheme use among learners with different L1s (Goldschneider and DeKeyser, 2001; Luk and Shirai, 2009). For example, Murakami and Alexopoulou (2016) found the presence or absence of articles and the plural *-s* in a learner's L1 correlated with the learner's accurate use of the equivalent L2 English morpheme. This thesis complements the previous morpheme order studies by expanding the range of languages, the number of morphemes, and the span of proficiency levels considered. This expansion is enabled by the NLI task which can canvas large data sets, and represent thousands of learners from numerous L1 backgrounds. Our findings will demonstrate the extent of CLI on morphemes produced by learners with different language backgrounds. The findings will have repercussions on how we understand the interaction of two languages in a language learner's productions. Furthermore, we contribute to the NLI task by focusing on morphosyntactic linguistic features which have not been investigated as the sole predictors in previous work.

3 Methodology

3.1 Corpus

We collect data from the EF Cambridge Open Language Database (EFCamDat), a longitudinal corpus consisting of 1.8 million English learner texts submitted by 174,000 students enrolled in a virtual learning environment (Huang et al., 2017). The corpus spans sixteen proficiency levels aligned with standards such as the Common European Framework of Reference for languages. The texts include metadata on the learner's proficiency level and nationality acts as a proxy to native language background. The texts are annotated with part of speech (POS) tags using the Penn Treebank Tagset, grammatical relationships using SyntaxNet, and some texts in the corpus (66%) include error annotations provided by teachers using predetermined error markup tags (Huang et al., 2017).

In addition to the English L2 subcorpus, we collect a corpus of English L1 writing samples. We crowdsource written responses through social media and undergraduate linguistics courses from self-described English monolingual speakers. Participants are asked to submit a paragraph addressing a prompt that mimics the EFCamDat tasks in the online English school. We collect enough responses to form an English L1 class that is similar in size and length to the English L2 subcorpus, which is described next.

3.2 Target Learner Groups

We create a subcorpus representing learners from ten L1 backgrounds: Arabic, Chinese, French, German, Japanese, Korean, Portuguese, Russian, Spanish, Turkish. We select these ten L1 backgrounds for three main reasons. First, typologically similar languages such as Spanish and Portuguese are included because more overt transfer effects occur in production when structural and functional cross-linguistic similarity is high (Ringbom, 2007). Thus we expect similar transfer effects that could make classification difficult between similar languages. Second, languages different from English in respect to orthography (e.g., Arabic) or word order (e.g., Korean and Russian) are included to detect if negative transfer, or erroneous use, occurs. Third, the ten selected groups

are represented in the TOEFL11 corpus which will be used for a cross-corpus evaluation.

The individual learner texts from each language background will be grouped by proficiency level. Since not all learners progress through all sixteen instructional units and some proficiency levels are overrepresented, the proficiency levels will be merged into four groups covering Levels 1-3, 4-7, 8-11, 12-16. Previous research shows clustering these proficiency levels has been useful for identifying the native language of learners (Jiang et al., 2014). We ensure text size is homogeneous by merging texts into 1,000 word tokens and compiling texts into equally-sized sets for each language background at each proficiency level. The texts retain information on individual identification numbers and writing topics.

3.3 Target Morphemes and Feature Set

The target morphemes include nine morphemes that frequently appear in the morpheme order studies: articles *a/an/the*, auxiliary *be*, copula *be*, plural *-s*, possessive *'s*, progressive *-ing*, irregular past, regular past tense *-ed*, and third person singular *-s*. We include short plural *-s* (e.g., *boot/boots*) and long plural *-es* (e.g., *hoof/hooves*) and exclude irregular plurals (e.g., *foot/feet*). The copula and auxiliary will include the first, second, and third person present and progressive forms, respectively. We make a distinction between the definite (*the*) and indefinite (*a/an*) articles because previous research suggests learners acquire definite and indefinite articles at different rates depending on their L1 background (Crosthwaite, 2016; Díez-Bedmar and Papp, 2008). The irregular and regular past tense forms will be limited to lexical verbs (e.g., *went, walked*) and will exclude modals (e.g., *would*) and passive voice (e.g., *stolen*). The third person singular form *-s* will include the allomorphs *-s* and *-es* (e.g., *she waits and watches*). The possessive *'s* will include forms attached to a noun phrase (e.g., *cat's tail* or *cats' tails*).

We use features that capture morphemes and morphosyntactic relationships. The feature set includes function words, lexical n-grams, POS n-grams, dependency features, and error corrections. Function words include topic-independent grammatical lexical items, and will capture definite/indefinite articles and auxiliary verbs. We use lexical words to capture regular and irregular past tense verbs. To avoid some topic bias, we remove proper nouns from the texts. We use POS 1-3grams and dependency features to capture morphosyntactic relationships such as possession, progressive *-ing*, plural *-s*, and third person singular *-s*. Error corrections provided by the EFCamDat metadata capture morpheme errors such as article misuse and incorrect plurals. We evaluate the predictive power of the features through a step-wise procedure.

4 Native Language Identification Task

We use a Support Vector Machine (SVM) classifier to evaluate the data, and obtain the following metrics: precision, recall, accuracy, and the F-measure. We use a one-vs-all approach, and evaluate each feature type using 10-fold cross-validation. The linear SVM classifier is selected given its sustained success in NLI work, as seen in the recent 2017 NLI shared task (Malmasi et al., 2017).

To evaluate feature information as it relates to morpheme acquisition, we compare the overuse, underuse, and erroneous use of the linguistic features within the ten L1 groups and native English group. Following a methodology proposed by (Malmasi and Dras, 2014), we use SVM weights to identify the most predictive features that are unique to each L1 class. We compare if the same best-performing features appear for each proficiency band. If the features are different in each proficiency band, then we suspect these may be proficiency related effects and we follow with post hoc analysis.

Post hoc analysis is conducted to compare the morphemes appearing more or less regularly given a specific L1 background, and across proficiency levels. This method allows us to piece together a diachronic view of morpheme use. We determine if the linguistic features show evidence of language transfer based on statistical evidence of between-group differences and within-group similarities. We expect to see differential use of morphemes between L1 groups that may or may not have equivalent morphemes between the L1 and L2. For example, Russian L1 learners may underuse English articles because they are absent in the L1. On the other hand, Spanish L1 learners may overuse articles because Spanish articles follow a different semantic scheme, thus English articles may be oversupplied to inappropriate con-

texts.

We also determine how well the classifier performs with native English data to each L1 for each proficiency level. We expect the classifier will show more confusion for typologically similar languages to English, especially at higher proficiency levels where written productions may contain fewer L1 idiosyncrasies.

5 Cross-Corpus Evaluation

In this section, we address the following questions: (1) How well do the most discriminatory linguistic features used in the EFCamDat corpus perform on an independent corpus? (2) Which features are corpus-specific? (3) Which features best predict the L1 class when trained on the EFCamDat and tested on an independent corpus, and vice versa. The classifier trained on the EFCamDat subcorpus will be used in a cross-corpus evaluation to increase the reliability of the linguistic predictors as evidence for language transfer.

We train on the EFCamDat subcorpus and test on an independent corpus, the TOEFL11, and vice versa. The TOEFL11 corpus consists of 12,100 English essays written during high-stakes college entrance exams from learners across eleven L1 backgrounds (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish) and distributed across eight prompts and three essay score levels (low/medium/high). The TOEFL11 corpus was designed with a native language identification task in mind and used in the first NLI shared task (Blanchard et al., 2013).

We match a subset of the learner groups from the TOEFL11 data with classes in the EFCamDat, use comparable text lengths for each language background, and match the proficiency levels between the two corpora. Success is measured by NLI accuracy that is higher than a chance baseline when the SVM is trained with the EFCamDat and tested on the TOEFL11, and vice versa. Most importantly, the corpus comparison determines which specific features serve as the strongest determiners of L1 background across proficiency levels, despite genre differences between the two corpora.

A cross-corpus methodology has the advantage of providing robust results that bolster the argument for CLI in the course of L2 morpheme development. The features that successfully distinguish between L1 groups will be analyzed to inform SLA research as to how certain morphemes may be more susceptible to language transfer than others, and how different L1 groups may follow unique trajectories of morpheme acquisition. Previous research has found some corpus-independent L1 transfer features that generalize across the different task requirements represented in the EFCamDat and TOEFL11 corpora (Malmasi and Dras, 2015). However, it is unknown if this generalizability will hold between proficiency bands. Thus, we test the classifier on different proficiency levels across EFCamDat and TOEFL11 to determine the features that are corpus-independent across proficiency levels. These findings will then be connected to formulating hypotheses on language transfer during the developmental trajectory of morpheme acquisition.

6 Human Cross-Validation

In this section, we address the following questions: (1) How does the performance of the classifier compare to humans performing the same task? (2) What linguistic predictors do humans use in classifying texts? To address these questions, we recruit native and non-native English speakers with a background in language instruction and/or self-reported linguistic training to perform a simplified online NLI task. We follow a similar procedure from Malmasi et al. (2015). The raters deal with five L1 classes representing the most disparate L1s from the EfCamDat subcorpus in order to facilitate classification. We split the texts into equal distributions of low and high proficiency essays for each L1 group. Raters perform the NLI task on roughly 50 essays and indicate the linguistic features that led to their decision and their confidence in that decision. We determine accuracy scores for the L1 groups across proficiency levels, and if the raters use morphemes and morphosyntactic relationships as indicators of the L1. The results of the study will indicate the qualitative and quantitative differences in detecting cross-linguistic influence based on human evaluations versus computational measures.

7 Conclusion

This thesis expands on NLI methodology and connects computational linguistics with SLA research. In terms of methodology, we investigate NLI using only grammatical morphemes, which have not been singled out in previous NLI re-

search. English is considered a morphologically weak language with comparatively few requirements for agreement and declensions. Achieving an accuracy score higher than chance indicates that morphemes, however ubiquitous in writing, can reveal a significant distribution that correctly identifies a writer's L1 background. The thesis may provide motivation to perform NLI on morphologically rich languages such as Slavic or Bantu, to identify if a classifier can use a wider set of morphemes as the sole feature. Furthermore, we expand the methodology to cross-corpus evaluations on four proficiency levels. This line of research shows how robust a model may be for lower to higher proficiency English learners.

In terms of combining computational linguistics with SLA, we use the NLI task as a tool for uncovering cross-linguistic influence that may otherwise go unseen by a human researcher. We test if that is indeed the case in the human cross-validation study. The NLI task increases the number and type of comparisons we can make between languages, which can lead to new insights in the underuse, overuse, and erroneous use of morphemes. We determine how learner groups develop the capacity to use morphemes through development. The automatic detection of transfer is especially valuable for higher proficiency learners, where transfer is harder to discern because the effects may not be obviously visible as errors (Ringbom, 2007). The ability to detect subtle CLI effects holds the potential to generate new hypotheses about where and why language transfer occurs, so that understanding can be transferred to L2 education. This thesis accounts for these transfer effects and provides testable hypotheses.

Acknowledgements

I am grateful to Dr. Stefanie Wulff for constant support and helpful discussions, and thank the anonymous reviewers for comments and suggestions on improving the thesis proposal.

References

Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. 2017. Predicting native language from gaze. *arXiv preprint arXiv:1704.07398*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2).

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. *Proceedings of COLING 2012*, pages 391–408.

Scott Crossley and Danielle McNamara. 2012. Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. *Approaching language transfer through text classification: Explorations in the detection-based approach*, pages 106–126.

Peter Crosthwaite. 2016. L2 english article use by l1 speakers of article-less languages. *International Journal of Learner Corpus Research*, 2(1):68–100.

María Belén Díez-Bedmar and Szilvia Papp. 2008. The use of the english article system by chinese and spanish learners. *Language and Computers Studies in Practical Linguistics*, 66:147.

Heidi Dulay and Marina Burt. 1974. Natural sequences in child second language acquisition. *Language Learning*, 24(1):37–53.

Jennifer M Goldschneider and Robert M DeKeyser. 2001. Explaining the natural order of l2 morpheme acquisition in english: A meta-analysis of multiple determinants. *Language Learning*, 51(1):1–50.

Yan Huang, Akira Murakami, Theodora Alexopoulou, and Anna Korhonen. 2017. Dependency parsing of learner english.

Xiao Jiang, Yufan Guo, Jeroen Geertzen, Dora Alexopoulou, Lin Sun, and Anna Korhonen. 2014. Native language identification using large, longitudinal data. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 3309–3312.

Stephen Krashen. 1985. *The input hypothesis: Issues and implications*. Addison-Wesley Longman Ltd.

Diane Larsen-Freeman. 1975. The acquisition of grammatical morphemes by adult esl students. *TESOL Quarterly*, pages 409–419.

Zoe Pei-sui Luk and Yasuhiro Shirai. 2009. Is the acquisition order of grammatical morphemes impervious to l1 knowledge? evidence from the acquisition of plural -s, articles, and possessive 's. *Language Learning*, 59(4):721–754.

Shervin Malmasi. 2016. *Native language identification: explorations and applications*. Ph.D. thesis, Macquarie University.

Shervin Malmasi and Mark Dras. 2014. Language transfer hypotheses with linear svm weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1385–1390.

Shervin Malmasi and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1403–1409.

Shervin Malmasi, 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 Building Educational Applications Using NLP*, Copenhagen, Denmark. Association for Computational Linguistics.

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015. Oracle and human baselines for native language identification. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–178.

Akira Murakami and Theodora Alexopoulou. 2016. L1 influence on the acquisition order of english grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38(3):365–401.

Håkan Ringbom. 2007. *Cross-linguistic similarity in foreign language learning*. Multilingual Matters.

Sabrina Stehwien and Sebastian Padó. 2015. Native language identification across text types: how special are scientists? *Italian Journal of Computational Linguistics*, 1(1).

Marcos Zampieri, Alina Maria Ciobanu, and Liviu P Dinu. 2017. Native language identification on text and speech. *arXiv preprint arXiv:1707.07182*.