Language Transfer Hypotheses with Linear SVM Weights

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

Language transfer, the characteristic second language usage patterns caused by native language interference, is investigated by Second Language Acquisition (SLA) researchers seeking to find overused and underused linguistic features. In this paper we develop and present a methodology for deriving ranked lists of such features. Using very large learner data, we show our method's ability to find relevant candidates using sophisticated linguistic features. To illustrate its applicability to SLA research, we formulate plausible language transfer hypotheses supported by current evidence. This is the first work to extend Native Language Identification to a broader linguistic interpretation of learner data and address the automatic extraction of underused features on a per-native language basis.

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.

Motivated by the aim of improving foreign language teaching and learning, such analyses are often done manually in SLA, and are difficult to perform for large corpora. Smaller studies yield poor results due to the sample size, leading to extreme variability (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 (L2) (Tetreault et al., 2013).

Despite the good results in predicting L1s, few attempts have been made to interpret the features that distinguish L1s. This is partly because no methods for an SLA-oriented feature analysis have been proposed; most work focuses on testing feature types using standard machine learning tools.

The overarching contribution of this work is to develop a methodology that enables the transformation of the NLI paradigm into SLA applications that can be used to link these features to their underlying linguistic causes and explanations. These candidates can then be applied in other areas such as remedial SLA strategies or error detection.

2 Related Work

SLA research aims to find distributional differences in language use between L1s, often referred to as **overuse**, the extensive use of some linguistic structures, and **underuse**, the underutilization of particular structures, also known as *avoidance* (Gass and Selinker, 2008). While there have been some attempts in SLA to use computational approaches on small-scale data,[1] these still use fairly elementary techniques and have several shortcomings, including in the manual approaches to annotation and the computational artefacts derived from these.

Conversely, NLI work has focused on automatic learner L1 classification using Machine Learning with large-scale data and sophisticated linguistic features (Tetreault et al., 2012). Here, feature ranking could be performed with relevancy methods such as the F-score:

[1] E.g. Chen (2013), Lozanó and Mendikoetxea (2010) and Diéz-Bedmar and Papp (2008).

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1385–1390,
October 25-29, 2014, Doha, Qatar.

$$F(j) \equiv \frac{\left(\bar{x}_j^{(+)} - \bar{x}_j\right)^2 + \left(\bar{x}_j^{(-)} - \bar{x}_j\right)^2}{\frac{1}{n_+-1}\sum_{i=1}^{n_+}\left(x_{i,j}^{(+)} - \bar{x}_j^{(+)}\right)^2 + \frac{1}{n_--1}\sum_{i=1}^{n_-}\left(x_{i,j}^{(-)} - \bar{x}_j^{(-)}\right)^2} \quad (1)$$

The F-score (Fisher score) measures the ratio between the intraclass and interclass variance in the values of feature j, where x represents the feature values in the negative and positive examples.[2] More discriminative features have higher scores.

Another alternative method is Information Gain (Yang and Pedersen, 1997). As defined in equation (2), it measures the entropy gain associated with feature t in assigning the class label c.

$$\begin{aligned} G(t) \;=\; & -\; \textstyle\sum_{i=1}^{m} \Pr(c_i) \log \Pr(c_i) \\ & +\; \Pr(t) \textstyle\sum_{i=1}^{m} \Pr(c_i|t) \log \Pr(c_i|t) \\ & +\; \Pr(\bar{t}) \textstyle\sum_{i=1}^{m} \Pr(c_i|\bar{t}) \log \Pr(c_i|\bar{t}) \end{aligned} \quad (2)$$

However, these methods are limited: they do not provide ranked lists per-L1 class, and more importantly, they do not explicitly capture underuse.

Among the efflorescence of NLI work, a new trend explored by Swanson and Charniak (2014) aims to extract lists of candidate language transfer features by comparing L2 data against the writer's L1 to find features where the L1 use is mirrored in L2 use. This allows the detection of obvious effects, but Jarvis and Crossley (2012) note (p. 183) that many transfer effects are "too complex" to observe in this manner. Moreover, this method is unable to detect underuse, is only suitable for syntactic features, and has only been applied to very small data (4,000 sentences) over three L1s. Addressing these issues is the focus of the present work.

3 Experimental Setup

3.1 Corpus

We use TOEFL11, the largest publicly available corpus of English L2 texts (Blanchard et al., 2013), containing 11 L1s with 1,100 texts each.[3]

3.2 Features

Adaptor grammar collocations Per Wong et al. (2012), we utilize an adaptor grammar to discover arbitrary length n-gram collocations. We explore both the pure part-of-speech (POS) n-grams as well as the more promising mixtures of POS and function words. We derive two adaptor grammars where each is associated with a different set of vocabulary: either pure POS or the mixture of POS and function words. We use the grammar proposed by Johnson (2010) for capturing topical collocations:

$Sentence \rightarrow Doc_j$	$j \in 1, \ldots, m$
$Doc_j \rightarrow \_j$	$j \in 1, \ldots, m$
$Doc_j \rightarrow Doc_j\ Topic_i$	$i \in 1, \ldots, t;$
	$j \in 1, \ldots, m$
$\underline{Topic_i} \rightarrow Words$	$i \in 1, \ldots, t$
$\overline{Words} \rightarrow Word$	
$Words \rightarrow Words\ Word$	
$Word \rightarrow w$	$w \in V_{pos};$
	$w \in V_{pos+fw}$

V_{pos} contains 119 distinct POS tags based on the Brown tagset and V_{pos+fw} is extended with 398 function words. The number of topics t is set to 50. The inference algorithm for the adaptor grammars are based on the Markov Chain Monte Carlo technique made available by Johnson (2010).[4]

Stanford dependencies We use Stanford dependencies as a syntactic feature: for each text we extract all the basic dependencies returned by the Stanford Parser (de Marneffe et al., 2006). We then generate all the variations for each of the dependencies (grammatical relations) by substituting each lemma with its corresponding POS tag. For instance, a grammatical relation of `det(knowledge, the)` yields the following variations: `det(NN, the)`, `det(knowledge, DT)`, and `det(NN, DT)`.

Lexical features Content and function words are also considered as two feature types related to learner's vocabulary and spelling.

3.3 Extracting Linear SVM Feature Weights

Using the extracted features, we train linear Support Vector Machine (SVM) models for each L1. We use a one-vs-rest approach to find features most relevant to each native language. L2-regularization is applied to remove noisy features and reduce the size of the candidate feature list. More specifically, we employ the LIBLINEAR SVM package (Fan et al., 2008)[5] as it is well-suited to text classification tasks with large numbers of features and texts as is the case here.

[2] See Chang and Lin (2008) for more details.

[3] Over 4 million tokens in 12,100 texts.

[4] http://web.science.mq.edu.au/%7Emjohnson/Software.htm

[5] http://www.csie.ntu.edu.tw/%7Ecjlin/liblinear/

In training the models for each feature, the SVM weight vector[6] is calculated according to (3):

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \quad (3)$$

After training, the positive and negative weights are split into two lists and ranked by weight. The positive weights represent overused features, while features whose absence (i.e. underuse) is indicative of an L1 class will have large negative weights. This yields two candidate language transfer feature lists per L1.

4 Results

We now turn to an analysis of the output from our system to illustrate its applicability for SLA research. Table 1 lists some elements from the underuse and overuse lists for various L1s. The lists are of different feature types. They have been chosen to demonstrate all feature types and also a variety of different languages. For reasons of space, only several of the top features are analysed here.

Hindi L1 writers are distinguished by certain function words including *hence*, *thus*, and *etc*, and a much higher usage rate of male pronouns. It has been observed in the literature (Sanyal, 2007, for example) that the English spoken in India still retains characteristics of the English that was spoken during the time of the Raj and the East India Company that have disappeared from other English varieties, so it sounds more formal to other speakers, or retains traces of an archaic business correspondence style; the features noted fit that pattern.

The second list includes content words overused by Arabic L1 learners. Analysis of content words here, and for other L1s in our data, reveals very frequent misspellings which are believed to be due to orthographic or phonetic influences (Tsur and Rappoport, 2007; Odlin, 1989). Since Arabic does not share orthography with English, we believe most of these are due to phonetics. Looking at items 1, 3 and 5 we can see a common pattern: the English letter *u* which has various phonetic realizations is being replaced by a vowel that more often represents that sound. Items 2 and 5 are also phonetically similar to the intended words.

For Spanish L1 authors we provide both underuse and overuse lists of syntactic dependencies. The top 3 overuse rules show the word *that* is very often used as the subject of verbs. This is almost certainly a consequence of the prominent syntactic role played by the Spanish word *que* which, depending on the context, is equivalent to the English words *whom*, *who*, *which*, and most commonly, *that*. The fourth rule shows they often use *this* as a determiner for plural nouns. A survey of the corpus reveals many such errors in texts of Spanish learners, e.g. *this actions* or *this emissions*. The fifth rule shows that the adjectival modifier of a plural noun is often being incorrectly pluralised to match the noun in number as would be required in Spanish, for example, *differents subjects*.

Turning to the underused features in Spanish L1 texts, we see that four related features rank highly, showing that *these* is not commonly used as a determiner for plural nouns and *which* is rarely used as a subject. The final feature shows that *no* is avoided as a determiner. This may be because while *no* mostly has the same role in Spanish as it does in English, it cannot be used as a determiner; *ningún* must be used instead. We hypothesize that this construction is being avoided as placing *no* before a noun in Spanish is ungrammatical. This example demonstrates that our two list methodology can not only help identify overused structures, but also uncovers the related constructs that are being underutilized at their expense.

The final list in Table 1 is of underused Adaptor Grammar patterns by Chinese learners. The first three features show that these writers significantly underuse determiners, here *an*, *other* and *these* before nouns. This is not unexpected since Chinese learners' difficulties with English articles are well known (Robertson, 2000). More interestingly, we find underuse of features like *even if* and *might*, along with others not listed here such as *could* VB[7] plus many other variants related to the subjunctive mood. One explanation is that linguistic differences between Chinese and English in expressing counterfactuals could cause them to avoid such constructions in L2 English. Previous research in this area has linked the absence of subjunctive linguistic structures in Chinese to different cognitive representations of the world and consequences for thinking counterfactually (Bloom, 2014), although this has been disputed (Au, 1983; Garbern Liu, 1985).

Adaptor Grammars also reveal frequent use of the "existential there"[8] in German L1 data while

[6] See Burges (1998) for a detailed explanation.

[7] e.g. *could be*, *could have*, *could go* and other variants

[8] e.g. *There is/are ...*, as opposed to the locative *there*.

Overuse			Underuse	
Hindi	Arabic	Spanish	Spanish	Chinese
#2: thus	#2: anderstand	#1: nsubj(VBP,that)	#2: det(NNS,these)	#12: *an* NN
#4: hence	#4: mony	#2: nsubj(VBZ,that)	#3: nsubj(VBZ,which)	#16: *other* NN
#22: his	#6: besy	#3: nsubj(VB,that)	#6: nsubj(VB,which)	#18: *these* NNS
#30: etc	#15: diffrent	#4: det(NNS,this)	#7: nsubj(VBP,which)	#19: *even if*
#33:rather	#38: seccessful	#25: amod(NNS,differents)	#10: det(NN,no)	#68: *might*

Table 1: Example transfer candidates and rankings from the overuse/underuse lists for various L1s and features types, in order: Hindi function words, Arabic content words, Spanish dependencies (2) and Chinese Adaptor Grammars.

English	**Spanish**	**English**	**Spanish**
diferent	diferente	conclution	conclusión
consecuence	consecuencia	desagree	Neg. affix *des-*
responsability	responsabilidad	especific	específico
oportunity	oportunidad	necesary	necesario

Table 2: Highly ranked English misspellings of Spanish learners and their Spanish cognates.

they are highly underused in French L1 data. The literature supports our data: The German equivalent *es gibt* is common while French use is far more constrained (Cappelle and Loock, 2013).

Lexical analysis also revealed Spanish–English orthographic transfer, listed in Table 2. This list includes many cognates, in contrast with the Arabic L1 data where most misspellings were phonetic in nature.

We also observe other patterns which remain unexplained. For instance, Chinese, Japanese and Korean speakers make excessive use of phrases such as *however*, *first* and *second*. One possibility is that this relates to argumentation styles that are possibly influenced by cultural norms. More broadly, this effect could also be teaching rather than transfer related. For example, it may be case that a widely-used text book for learning English in Korea happens to overuse this construction.

Some recent findings from the 2013 NLI Shared Task found that L1 Hindi and Telugu learners of English had similar transfer effects and their writings were difficult to distinguish. It has been posited that this is likely due to shared culture and teaching environments (Malmasi et al., 2013).

Despite some clearcut instances of overuse,[9] more research is required to determine the causal factors. We hope to expand on this in future work using more data.

[9]More than half of the Korean scripts contained a sentence-initial *however*.

5 Discussion and Conclusion

Using the proposed methodology, we generated lists of linguistic features overused and underused by English learners of various L1 backgrounds. Through an analysis of the top items in these ranked lists, we demonstrated the high applicability of the output by formulating plausible language transfer hypotheses supported by current evidence. We also showcased the method's generalizability to numerous linguistic feature types.

Our method's output consists of two ranked lists of linguistic features: one for overuse and the other for underuse, something which had not been addressed by research to date. We also found Adaptor Grammar collocations to be highly informative for this task.

This work, an intersection of NLP, Machine Learning and SLA, illustrates how the various disciplines can complement each other by bringing together theoretical, experimental and computational issues. NLP provides accurate and automated tagging of large corpora with sophisticated features not available in corpus linguistics, e.g. with state-of-the-art dependency parsing. Sophisticated machine learning techniques then enable the processing of large quantities of data (thousands of times the size of manual studies) in a way that will let SLA researchers explore a variety of assumptions and theoretical analyses. And conversely, NLP can benefit from the long-term study and language acquisition insights from SLA.

In terms of NLI, this work is the first attempt to expand NLI to a broad linguistic interpretation of the data, including feature underuse. NLI systems achieve classification accuracies of over 80% on this 11-class task, leading to theoretical questions about the features that make them so effective. This work also has a backwards link in this regard by providing qualitative evidence about the underpinning linguistic theories that make NLI work.

The work presented here has a number of applications; chief among them is the development of tools for SLA researchers. This would enable them to not just provide new evidence for previous findings, but to also perform semi-automated data-driven generation of new and viable hypotheses. This, in turn, can help reduce expert effort and involvement in the process, particularly as such studies expand to more corpora and emerging language like Chinese (Malmasi and Dras, 2014b) and Arabic (Malmasi and Dras, 2014a).

The brief analysis included here represents only a tiny portion of what can be achieved with this methodology. We included but a few of the thousands of features revealed by this method; practical SLA tools based on this would have a great impact on current research.

In addition to language transfer hypotheses, such systems could also be applied to aid development of pedagogical material within a needs-based and data-driven approach. Once language use patterns are uncovered, they can be assessed for teachability and used to create tailored, L1-specific exercises and teaching material.

From the examples discussed in Section 4 these could include highly specific and targeted student exercises to improve spelling, expand vocabulary and enrich syntactic knowledge — all relative to their mother tongue. Such exercises can not only help beginners improve their fundamental skills and redress their errors but also assist advanced learners in moving closer to near-nativeness.

The extracted features and their weights could also be used to build statistical models for grammatical error detection (Leacock et al., 2014). Contrary to the norm of developing error checkers for native writers, such models could be specifically targeted towards learners or even particular L1–L2 pairs which could be useful in Computer-Assisted Language Learning (CALL) systems.

One limitation here is that our features may be corpus-dependent as they are all exam essays. This can be addressed by augmenting the data with new learner corpora, as they become available. While a strength here is that we compared each L1 against others, a paired comparison only against native texts can be insightful too.

There are several directions for future work. The first relates to clustering the data within the lists. Our intuition is that there might be coherent clusters of related features, with these clusters characterising typical errors or idiosyncrasies, that are predictive of a particular L1. As shown in our results, some features are highly related and may be caused by the same underlying transfer phenomena. For example, our list of overused syntactic constructs by Spanish learners includes three high ranking features related to the same transfer effect. The use of unsupervised learning methods such as Bayesian mixture models may be appropriate here. For parse features, tree kernels could help measure similarity between the trees and fragments (Collins and Duffy, 2001).

Another avenue is to implement weight-based ranking methods to further refine and re-rank the lists, potentially by incorporating the measures mentioned in Section 2 to assign weights to features. As the corpus we used includes learner proficiency metadata, it may also be possible to create proficiency-segregated models to find the features that characterise errors at each language proficiency level. Finally, the use of other linguistic features such as Context-free Grammar phrase structure rules or Tree Substitution Grammars could provide additional insights.

In addition to these further technical investigations, we see as a particularly useful direction the development of an SLA research tool to conduct a large SLA study with a wide range of experts. We believe that this study makes a contribution to this area and hope that it will motivate future work.

References

Terry Kit-Fong Au. 1983. Chinese and English counterfactuals: the Sapir-Whorf hypothesis revisited. *Cognition*, 15(1):155–187.

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

Alfred H Bloom. 2014. *The linguistic shaping of thought: A study in the impact of language on thinking in China and the West*. Psychology Press.

Christopher JC Burges. 1998. A tutorial on Support Vector Machines for Pattern Recognition. *Data mining and knowledge discovery*, 2(2):121–167.

Bert Cappelle and Rudy Loock. 2013. Is there interference of usage constraints?: A frequency study of existential there is and its French equivalent il ya in translated vs. non-translated texts. *Target*, 25(2):252–275.

Yin-Wen Chang and Chih-Jen Lin. 2008. Feature ranking using linear svm. *Causation and Prediction Challenge Challenges in Machine Learning, Volume 2*, page 47.

Meilin Chen. 2013. Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics*, 18(3).

Michael Collins and Nigel Duffy. 2001. Convolution Kernels for Natural Language. In *Advances in Neural Information Processing Systems*, pages 625–632.

Marie-Catherine de Marneffe, Bill Maccartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 449–454, Genoa, Italy.

María Belén Diéz-Bedmar and Szilvia Papp. 2008. The use of the English article system by Chinese and Spanish learners. *Language and Computers*, 66(1):147–176.

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.

Lisa Garbern Liu. 1985. Reasoning counterfactually in Chinese: Are there any obstacles? *Cognition*, 21(3):239–270.

Susan M. Gass and Larry Selinker. 2008. *Second Language Acquisition: An Introductory Course*. Routledge, New York.

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

Mark Johnson. 2010. PCFGs, Topic Models, Adaptor Grammars and Learning Topical Collocations and the Structure of Proper Names. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1157, Uppsala, Sweden, July. Association for Computational Linguistics.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies*, 7(1):1–170.

Cristobal Lozanó and Amaya Mendikoetxea. 2010. Interface conditions on postverbal subjects: A corpus study of L2 English. *Bilingualism: Language and Cognition*, 13(4):475–497.

Shervin Malmasi and Mark Dras. 2014a. Arabic Native Language Identification. In *Proceedings of the Arabic Natural Language Processing Workshop (co-located with EMNLP 2014)*, 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, 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.

Daniel Robertson. 2000. Variability in the use of the English article system by Chinese learners of English. *Second Language Research*, 16(2):135–172.

Jyoti Sanyal. 2007. *Indlish: The Book for Every English-Speaking Indian*. Viva Books Private Limited.

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

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.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2012. Exploring Adaptor Grammars for Native Language Identification. In *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pages 699–709.

Yiming Yang and Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420.