Native Language Identification Using Collocations as Features

Haiyin Yang
Department of Linguistics
University of Florida
haiyin.yang@ufl.edu

1 Introduction

Native language identification (NLI) aims to identify the L1 of a writer based on their L2 writing. The task is particularly interesting to the study of second language acquisition because it reveals important transfer patterns between L1s and L2s. Previous NLI studies used syntactical structures (e.g. part-of-speech and dependency tags) and lexical features (word and character n-grams) for NLI tasks (Berzak et al., 2014; Gyawali et al.; Liu et al., 2022). Though these features may achieve high accuracy, they are hard to interpret and thus could not reveal much about the exact syntactical or lexical structures transferred from different L1s. On the other hand, second language acquisition research has shown that the L1 also impacts the L2 lexicon in various ways: L1 interference manifests in the collocations that learners produce (Laufer and Waldman, 2011; Paquot, 2013; Wu and Tissari, 2021), and L2 learners bootstrap word chunks to yield an increasingly productive collocation repertoire (Ellis, 1996, 2012). At the same time, collocation frequencies affect native speakers' perception (Hilpert, 2008), processing (Kapatsinski and Radicke, 2009), and priming effects (Durrant and Doherty, 2010). Unlike n-grams that cut phrase boundaries, collocations are units of formulaic language revealing psychological associations between words in the mental lexicon (Hoey, 2005). Thus, collocations deserve more attention in NLI research than they have been given so far. This research intends to address this gap by leveraging collocations as classification features to investigate whether collocations are effective NLI features. Our positive result suggests that this method can be applied to large-scale data to reveal cross-linguistic collocation transfer patterns and provide candidates for collocation transfer between understudied L1/L2 pairs.

2 Method

2.1 Data cleaning and collocation structures

We used the International Corpus of Learner English (ICLE) (Granger et al., 2020) and its native writing counterpart, The Louvian Corpus of Native English Essays (LOCNESS) (Granger, 1998). The programming language we used is Python (Python Software Foundation, https://www.python.org/)). We deleted L1s with sample sizes fewer than two percent of the whole data, leaving 16 L1s (Russian, Finnish, Spanish, Czech, Norwegian, Chinese, Turkish, Japanese, French, Bulgarian, Italian, Tswana, Swedish, Polish, German, and British and American). The collocation features' structures, categories, and lengths are adopted from previous L2 collocation studies. Four structures of collocations have been used: 1) adverb-verb pairs (Wu and Tissari, 2021), 2) a three-word bundle with a verb (Paquot, 2013), 3) verb-noun pairs (Nesselhauf, 2003), and 4) adjective-noun pairs (Siyanova and Schmitt, 2008). Dependency parsing information, calculated using the python package Spacy (https://spacy.io/) is used to ensure that 1) the adverb is a child of the verb, the adjective the child of the noun, and the noun a child of the verb so that these form meaningful collocates, not just adjacent word bundles, 2) in the three-word bundle that contains a verb, the verb is a member of the ancestors of the two other words, so the three-word bundle does not spread across the clause whose root is the verb.

2.2 Feature reduction

To achieve a balance between number of features and model performance, we used the 10-fold cross-validation with the following steps: 1) We selected collocates used by at least n% of texts from an L1 group; 2) To ensure that the word bundles were used homogeneously in an L1 group and heterogeneously in some other L1 groups (Jarvis, 2000), we

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 365-369.
Irvine, California, June 27-29, 2024

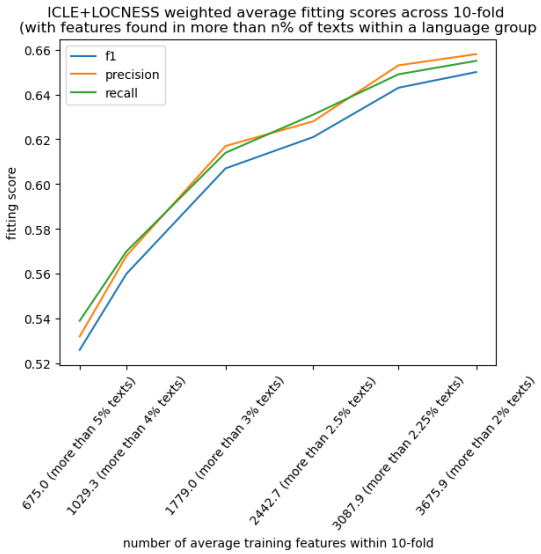

Figure 1: The number of training features v.s. f1, precision, and recall rates. The numbers are averaged across 10-fold validation.

applied one-way ANOVA tests (Paquot, 2013).

2.3 Classification

The Ridge Classifier from sklearn is used in this project because 1) it is time-efficient, 2) it avoids large coefficients, and 3) the coefficient levels can reveal the prediction power of features for each L1 group. Avoiding large coefficients is essential for this task because the project uses lexical features, and about 45% of features in the ICLE training set do not reappear in the test set. If some features have high coefficients but do not appear in the testing data, their prediction power is wasted.

3 Analysis

3.1 The prediction power of collocation features: sample size and collocation idiosyncrasies

The fitting scores of the model demonstrate that collocations provide prediction power in the NLI task. Figure 1 shows the f1, precision, and recall rates averaged across 10-fold validation with different numbers of features, demonstrating diminishing returns of features for fitting scores. The rate of increase drastically declining after around 3,100 features. A good balance between the number of features and performance is between 1,800 and 3,100 features. The rest of the analysis in this paper uses about 2,400 features with an f1 of 62%.

The precision rates vary across L1s, as shown in figure 2. One potential reason causing the lower fitting scores for some L1s is the unbalanced sample sizes. L1 groups with precision scores lower than 50% (German, Norwegian, Czech, Finnish, Swedish) all have below-average training data sizes. Their lower performances may thus be caused by insufficient training size. Moreover, as the L1 Chinese group contributes to a large portion of the data (16%), the classifier tends to misclassify other L1 groups as L1 Chinese to achieve a higher fitting score.

We also performed hierarchical clustering to visualize the degree of collocation production similarities among L1 groups so we could understand their interaction with fitting scores. For each L1 group, we counted the occurrences of collocates used by at least 2.5% of within-group samples and passing the ANOVA test, obtaining a vector documenting the hits of individual collocates for each L1. The vectors were then normalized and inputted into hierarchical clustering using Ward's algorithm (Ward, 1963), a bottom-up clustering method that minimizes within-cluster variance. We used the Python package sklearn (Pedregosa et al., 2011) to implement the clustering. The height of the horizontal branches where two clusters merge can be interpreted as a measure of their differences, and lower height implies higher similarity.

The groups with the highest precision rates are Chinese (89%), Tswana (77%), Japanese (78%), and Italian (68%). L1 Chinese and Tswana groups have larger sample sizes (16% and 8.5% of the total data), which may contribute to their high fitting scores. On the other hand, Japanese and Italian have average sample sizes (around 6%). The commonality among these four groups, however, is that their samples contain more idiosyncratic collocation features, as manifested by their high branch levels in figure 3. Therefore, collocation production idiosyncrasies of L1 groups affect the model performance.

3.2 Potential Production Similarities from Different L1s

To investigate the misclassification of the model and whether this reveals collocation production similarities between groups, we plotted a normalized confusion matrix (figure 4) that shows the percentages of predicted labels for each true label. Each row adds to 100%. For instance, the second cell of the first row is 5.2%, which means that the

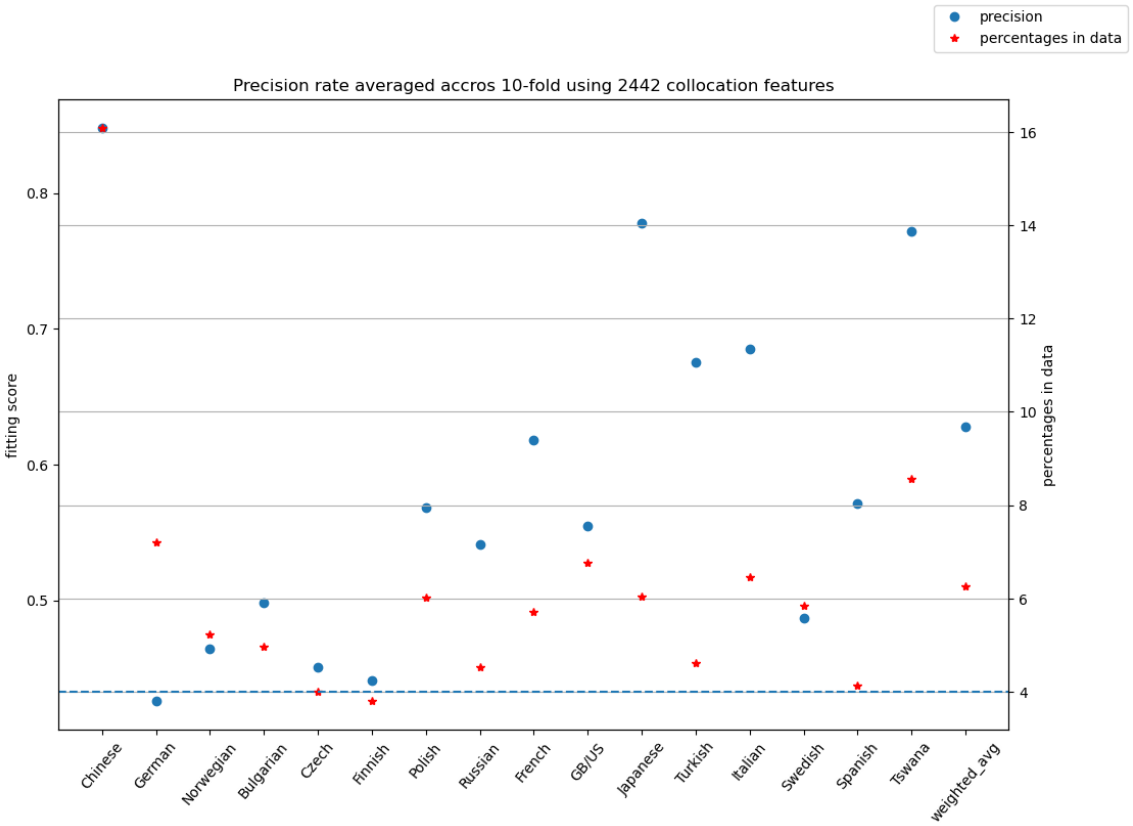

Figure 2: Sample sizes relative to total data and precision rates averaged across 10-fold validation using 2442 collocation features.

classifier misclassifies 5.2% of Bulgarian L1s as native writing. The confusion matrix aligns with the clustering dendrogram to some extent: 13.7% of Russian L1 tends to be misclassified as Bulgarian, whose collocation productions turn out to be close to those of Russian shown in the clustering dendrogram; Another small-distance cluster (in the middle of the dendrogram consisting of Swedish, German, Norwegian, Finnish, and native English) may explain the high percentage of Swedish L1 misclassified as German (12%), native English (6.7%), and Norwegian (6.7%), and the high percentage of German L1 misclassified as Swedish (7.8%).

3.3 Collocation features aligning with previous SLA studies

The high coefficient features are the signals the classifier identified for each L1. We compared the machine-identified features with available L2 collocation studies to see if the classifier is able to find valid collocation transfers.

The L1 groups we examined are French and Chinese, both with high classification results. The three-word collocations with high coefficients assigned by the L1 French classifier contain 9 of 15 collocations identified by Paquot (2013). Wu & Tissari (2021) found that Chinese learners of English produce much fewer types of intensifiers compared to English native speakers, and indeed, high-coefficient features for the L1 Chinese group contain much fewer intensifiers compared to those of native writers and the L1 French group (5 vs. 12 and 10). This reveals that the classifier can pick valid collocations, and the types of collocations reveal production patterns.

3.4 Removing native data

One potential concern with the dataset could be that it includes both L2 and L1 writings, though both are college students' essays. Therefore, we performed an analysis removing the L1 data to gauge its influence on the model. Compared to the full dataset, removing the L1 data makes a small impact on the fitting scores (mean f1 difference = 0.012, standard deviation of f1 difference = 0.031).

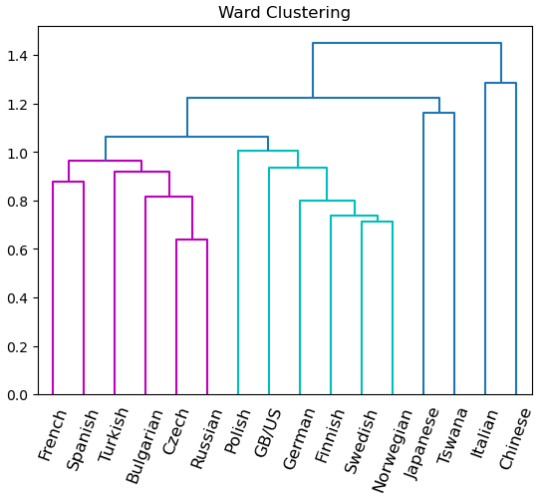

Figure 3: Hierarchical clustering dendrogram based on collocations of L1s. Ward's algorithm that minimizes within-cluster variance is used to compute the clustering.

We thus concluded that the inclusion of native data poses little impact on the results.

4 Limitation

Since this project utilizes lexical features, which tend to have fewer occurrences in test data, the sparse feature matrix in the test data harms the fitting scores. Longer texts may allow more hits of the features and thus improve model performance.

Although one contribution of this project is identifying potential collocation transfer, the probability of these collocations as real transfer depends partially on the classifier's performance. For L1s with high fitting scores, such as Chinese, Japanese, and Tswana, the confidence that their features are collocation transfers is high. However, for L1s with low classification performance, such as Czech and Finnish, the features selected by the classifier may have less value for transfer identification. A more balanced training sample is likely to improve the collocation validity for more languages.

The model did not test the potential impact of topics on the collocation productions. An L1 group with a predominant topic that elicits unique collocations from this L1 group could cause high model performance for this L1, while the model features would not reveal meaningful L1 transfer. Future investigation is needed to gauge the influence of topics on model performance.

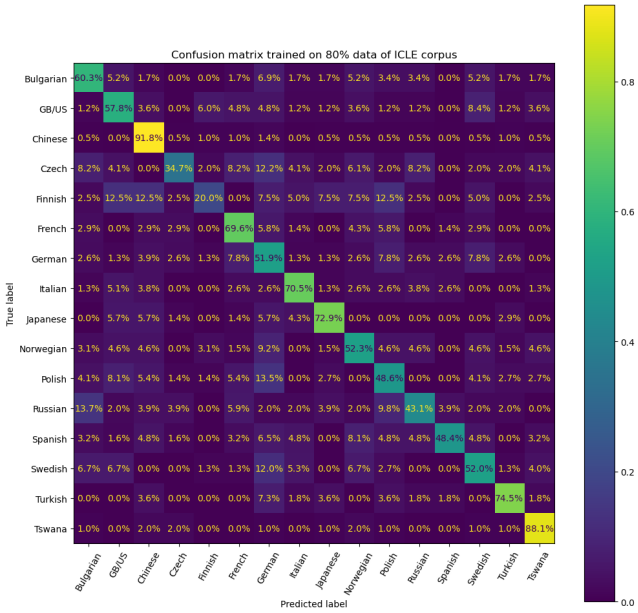

Figure 4: Confusion matrix of the ridge classifier with a training size of 80%. The summation of each row is 100%.

4.1 Conclusion and future direction

This project reveals that collocations have prediction power for NLI, and the results provide insights into collocation transfer. Specifically, it shows that this method can 1) provide candidates for collocation transfer, as those with high positive coefficients are vital signals for the corresponding L1; 2) reveal common patterns of how learners from different L1s produce English; and 3) reveal similarities and idiosyncracies in L2 collocation productions across different L1s.

Future research on large-scale corpora with different L1/L2 pairs can answer research questions concerning crosslinguistic collocation transfers, such as the types of collocate structures that tend to transfer across different language pairs.

Acknowledgments

This project would not have been possible without Dr. Zoey Liu's advice and Dr. Stefanie Wulff's inspiration. We are also thankful for the constructive feedback from the anonymous reviewers.

References

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. *Proceedings of the Eighteenth*

Conference on Computational Language Learning, pages 21–29.

Philip Durrant and Alice Doherty. 2010. Are high-frequency collocations psychologically real? investigating the thesis of collocational priming. *Corpus linguistics and linguistic theory*, 6(2):125–155.

Nick C. Ellis. 1996. Sequencing in lsa: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18:91–126.

Nick C. Ellis. 2012. Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, 32:17–44.

S. Granger. 1998. *The computer learner corpus: A versatile new source of data for SLA research*, pages 3–18. Addison Wesley Longman, London New York.

S. Granger, M. Dupont, F. Meunier, H. Naets, and M. Paquot. 2020. *The International Corpus of Learner English. Version 3.* Louvainla-Neuve: Presses universitaires de Louvain.

Binod Gyawali, Gabriela Ramirez, and Thamar Solorio. Native language identification: a simple n-gram based approach. Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 224–231. Association for Computational Linguistics.

Martin Hilpert. 2008. New evidence against the modularity of grammar: Constructions, collocations, and speech perception. *Cognitive linguistics*, 19(3):491–511.

Michael Hoey. 2005. *Lexical priming : a new theory of words and language*. Routledge, London, UK.

S. Jarvis. 2000. Methodological rigor in the study of transfer : Identifying l1 influence in the interlanguage lexicon. *Language learning*, 50(2):245–309.

Vsevolod Kapatsinski and Joshua Radicke. 2009. Frequency and the emergence of prefabs: Evidence from monitoring. *Formulaic language: Acquisition, loss, psychological reality, functional explanations*, 2:499–520.

Batia Laufer and Tina Waldman. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners' english. *Language learning*, 61(2):647–672.

Zoey Liu, Tiwalayo Eisape, Emily Prud'hommeaux, and J. K Hartshorne. 2022. Data-driven crosslinguistic syntactic transfer in second language learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44.

Nadja Nesselhauf. 2003. The use of collocations by advanced learners of english and some implications for teaching. *Applied linguistics*, 24(2):223–242.

Magali Paquot. 2013. Lexical bundles and l1 transfer effects. *International journal of corpus linguistics*, 18(3):391–417.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Anna Siyanova and Norbert Schmitt. 2008. L2 learner production and processing of collocation: A multi-study perspective. *Canadian modern language review*, 64(3):429–458.

Jr Ward, Joe H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Junyu Wu and Heli Tissari. 2021. Intensifier-verb collocations in academic english by chinese learners compared to native-speaker students. *Chinese journal of applied linguistics*, 44(4):470–487.