Using NLI to Identify Potential Collocation Transfer in L2 English

Haiyin Yang, Zoey Liu, Stefanie Wulff
University of Florida, FL, U.S.A.
{haiyin.yang, liu.ying, swulff} @ufl.edu

Abstract

Identifying instances of first language (L1) transfer – the application of the linguistics structures of a speaker's first language to their second language(s) – can facilitate second language (L2) learning as it can inform learning and teaching resources, especially when instances of negative transfer (that is, interference) can be identified. While studies of transfer between two languages A and B require a priori linguistic structures to be analyzed with three datasets (data from L1 speakers of language A, L1 speakers of language B, and L2 speakers of A or B), native language identification (NLI) – a machine learning task to predict one's L1 based on one's L2 production – has the advantage to detect instances of subtle and unpredicted transfer, casting a "wide net" to capture patterns of transfer that were missed before (Jarvis and Crossley, 2018). This study aims to apply NLI tasks to find potential instances of transfer of collocations. Our results, compared to previous transfer studies, indicate that NLI can be used to reveal collocation transfer, also in understudied L2 languages.

1 Introduction

The investigation of first language (L1) transfer is fascinating not only because it reveals how the brain processes two languages, but also because the identification of L1 transfer can help direct learning and teaching resources to areas where transfer, especially negative transfer, interferes with efficient communication. Corpus (learner production) data provide valuable insights into identifying instances of L1 transfer on L2 production. For L1 language A and L2 language B, transfer effect can be tested – given data of L1 speakers of A, L1 speakers of B, and L2 speakers of B – based on intragroup homogeneity (the distribution of the candidate of transfer need to be homogenous in this L1 group), intergroup heterogeneity (it is not the case that the distribution of the candidate of transfer is the same across all different backgrounds of L1s), and intra-L1-group congruity (the linguistic pattern of the candidate of transfer can be found in the native production of the L1 language) (Jarvis, 2000) to confirm that the proposed instances of linguistic structures come indeed from L1 transfer. The limitation of this approach is that 1) one needs to start with a priori linguistic structures to test, and 2) the L1 and L2 languages one can work with depend not only on available L2 data but also L1 data.

On the other hand, Native Language Identification (NLI) (Koppel et al., 2005; Malmasi and Dras, 2015; Markov et al., 2020; Ionescu and Popescu, 2017; Lotfi et al., 2020), a machine learning task that aims to identify the L1 of a language user based on their L2 production, is particularly applicable to the study of L2 learning because it can reveal transfer patterns between L1 and L2. Linguistic features that have high predictive power to identify the L1 background of a language producer can distinguish these speakers from those of other L1 backgrounds, i.e., features highly possible with intergroup heterogeneity and intragroup homogeneity. Therefore, NLI models can be used to identify potential instances of linguistic transfer (or transfer candidates) for multiple L1/L2 pairs.

This study aims to test the potential of leveraging NLI to find instances of transfer, and specifically, those of collocations (frequently co-occurring lexical combinations within a phrase). We focus on collocations for the following reasons. First, collocations are easily interpretable features. They are units of formulaic language that reveal psychological associations between words in the mental lexicon (Hoey, 2005). Compared to other common features of NLI tasks, such as syntactical structures (e.g., n-grams of part-of-speech tags and dependency tags) and pure lexical features that ignore word-dependency relationships (word and character n–grams), collocations features can be implemented in L2 pedagogy more straightforwardly.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 687–696
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Second, studies have found that second language learners tend to struggle with collocation acquisition (Nesselhauf, 2003; Laufer and Waldman, 2011), and L1 collocations interfere with L2 production (Paquot, 2013; Wu and Tissari, 2021). This may lead to communication inefficiency (e.g., the use of 'deliver a discussion' instead of 'hold a discussion'), and thus, identifying transfer of collocations can facilitate L2 production.

We ask the following research questions: 1) In this NLI task, do collocation features with high predictive power align with those identified for this specific L1/L2 pair in previous analyses? In other words, does the machine actually select those that are highly likely to be collocation transfer? 2) Why do we observe low performance for some L1s? In order to address the first question, we built a ridge classifier with collocations as features, selected two L1s, and compared the features with high coefficient values to the findings of previous transfer studies. To address the second question, we performed hierarchical clustering and compared it to the confusion matrix.

Testing on English L2 data (15 L1s, 5,600 pieces of writing), our positive NLI results suggest that this method can be used to cast a broad net to capture collocation transfer for multiple L1s, and specifically for understudied L2 languages.

2 Literature Review

2.1 Collocations and L1 transfer

Collocations, or words that often occur together within a phrase (Sinclair, 1991; Cowie, 2006), are units of formulaic language revealing psychological associations between words in the mental lexicon. Collocation frequencies affect native speakers' perception (Hilpert, 2008), processing (Kapatsinski and Radicke, 2009), and priming effects (Durrant and Doherty, 2010). These effects can be explained by the knowledge the mind has accumulated from the frequent association of a word. In other words, processing of a word primes the mind to activate words that frequently occur with it.

Moreover, research has shown that L1 collocation knowledge impacts L2 production (e.g., Laufer and Waldman 2011; Paquot 2013; Wu and Tissari 2021) and processing (e.g., Wolter and Gyllstad 2011; Cangır and Durrant 2021). For instance, Wu and Tissari (2021) found that Chinese learners of English use fewer types of intensifiers with verbs compared to native English writers, which can be explained by the fewer number of intensifiers in Chinese compared to English. Psycholinguistic tests also show that the L1 affects the processing of collocations in the L2. Wolter and Gyllstad (2011), using lexical decision task, found that, for Swedish learners of English, an L2 verb-noun collocation congruent with the L1 tends to be processed faster in general than an L2 collocation that has no translation equivalent in Swedish. Cangır and Durrant (2021), also using lexical decision task, even found cross-linguistic transfer effects in Turkish learners of English, who demonstrated positive priming effects with adjective-noun collocations when seeing the adjective in Turkish and the noun in English. These findings suggest that lexical knowledge of the L1 impacts both the production and processing of L2 collocations.

Besides the impact on production and processing, studies have also found that L2 learners tend to struggle with collocation acquisition. Focusing on verb-noun collocations produced by Hebrew learners of English, Laufer and Waldman (2011) found that learners underuse the collocations that native speakers frequently use, and L1 influence probably caused them to choose erroneous verb-noun combinations. Nesselhauf (2003) also found that learners have difficulty acquiring native-like L2 collocations: Using learner production from the German Corpus of Learner English (GeCLE), she found that more than half of the verb-noun collocations produced by German learners of English were erroneous or questionable.

2.2 Native language identification

The basic idea behind native language identification is that the native language impacts one's second language (Krashen, 1981), leaving "fingerprints" on L2 production. NLI can thus detect the linguistic features of transfer and the extent of transfer. Jarvis calls this a "detection-based approach", i.e., leveraging the intragroup homogeneity and intergroup heterogeneity, which signals group-based behavior that is distinct from other L1 groups, to capture linguistic transfer features (Jarvis and Crossley, 2018). Another method to identify linguistic transfer is the so-called "comparison-based approach", where one leverages statistical significance tests to find evidence from group-based behavior and rules out other factors that could potentially lead to its occurrence (i.e., topic, proficiency) using comparison to source-based behavior. Both approaches have different strengths: While the "comparison-based approach" is good at ruling out false-positive findings (i.e., identifying a feature as transfer while actually it is not), the "detection-based approach" excels in finding subtle, unpredicted, or indirect features of transfer that do not align with the L1 language (e.g., avoidance of certain structures, over corrections) (Jarvis and Crossley, 2018).

Frequent linguistic features used in NLI include lexical features (e.g., word frequencies and word n-grams) and syntactic features (e.g., dependency relationships n-grams, part-of-speech (POS) tag frequencies and POS n-grams) (see Goswami et al., 2024 for a review of NLI studies). While these studies focused on feature engineering and model performance, only a few (e.g., Liu et al., 2022) investigated the interpretability of these models or implications regarding cross-linguistic impact (Goswami et al., 2024). Because collocations are regarded as formulaic language expressions stored in one's language repertoire and hence readily interpretable, they are chosen as features in this study to showcase the potential of the NLI task as a tool to reveal language transfer patterns.

3 Method

3.1 Data

We use the International Corpus of Learner English (Granger et al., 2020), a corpus of college student essays, as the training and testing corpus. L1s whose number of essays is fewer than two percent of the whole data size are excluded, with 15 L1s (Russian, Finnish, Spanish, Czech, Norwegian, Chinese, Turkish, Japanese, French, Bulgarian, Italian, Tswana, Swedish, Polish, German) remaining in the study. The sample size of each L1 is unbalanced (mean = 379, standard deviation = 171), with L1 Chinese as the largest group (N = 980) contributing approximately 16% of the total sample size, and L1 Finnish as the smallest group (N = 230) contributing less than 4% of the total size. On average, each text is about 600 words.

The best clue for topic information of each essay is its prompt, which can be found from the ICLE metadata. In some L1 groups, each prompt is shared among tens to hundreds of essays (e.g., Bulgarian), while in others, a significant portion of the essays use idiosyncratic prompts. See Figure 1 for the frequency of prompts in each L1 group.

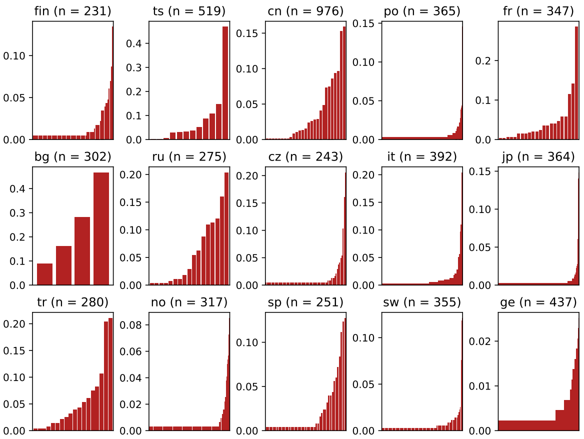


Figure 1: A histogram of the relative frequency of prompts in each L1 groups. Bars represent unique prompts, sorted by their relative frequencies in the L1 group.

3.2 Feature extraction, reduction, and topic-influence removal

The collocation features' structures, categories, and lengths are adopted from previous L2 collocation studies. Four structures of collocations are 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 (derived from the Python package *spaCy* Honnibal et al. 2020) is used to ensure that the extracted features are indeed collocations, not just neighboring words: 1) the adverb is a child of (i.e., modifies) the verb, the adjective is the child (i.e., modifies) the noun, and the noun is a child (i.e., an object) of the verb, 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 (for instance, in the sentence *"The unicorn who can fly, surprisingly, can also sing"*, *surprisingly* does not modify *fly*; if parsed correctly, *surprisingly* is not a child of *fly*, hence *can fly surprisingly* is not counted as a feature).

To achieve a balance between the number of features and model performance, and to address topic influence on lexical features, the following feature filtering steps are used together with 10-fold cross-validation. First, collocates used by at least n% of texts from an L1 group are selected as training features. To ensure that the word bundles were used homogeneously in an L1 group and heterogeneously in other L1 groups, one-way ANOVA test is applied to the lexical features (Paquot, 2013).

In order to address the topic's influence on lexical features, we approximated the dispersion of prompts where a feature appears via its entropy value. A collocation that is independent from topic influence is likely to appear in all prompts equally likely, and would thus have a high entropy value, whereas a collocation occurring due to topic influence would appear in limited prompts, resulting in a low entropy value. For a feature in an L1 group, its entropy value is calculated as Eq (1) below, where p_i is the estimated probability of $prompt_i$ from the pool of essays containing this feature, and T, the base of log, is the number of unique prompts in this L1 group. The base of log is set this way so that entropy values of features from L1s of different number of prompts can be fairly compared. An entropy value is always one if its probability to occur in each prompt is equal, regardless of how many prompts there are in the L1 group. Features with entropy values lower than 0.25 are removed. [1]

$$-\Sigma p_i \cdot log_T(p_i) \qquad (1)$$

Finally, 10-fold validation is used to obtain a reliable fitting result. Within each iteration, training features are reduced via steps outlined in the previous two paragraphs. The TfidfVectorizer function from the package *sklearn* (Buitinck et al., 2013), which counts the frequency of each feature in a text and weights a feature's text-wide frequency based on its corpus-wide frequency, is used with default parameters to compute the input matrix. For a feature, the smaller the corpus-wide frequency, the higher the weight. This is because if a feature is ubiquitous in the corpus and thus shared by many texts with different labels, it probably has low prediction power and thus receives a lower weight. After the feature counts are weighted, TfidVectorizer performs normalization so that the sum of squares of the feature frequency for one data point is 1.

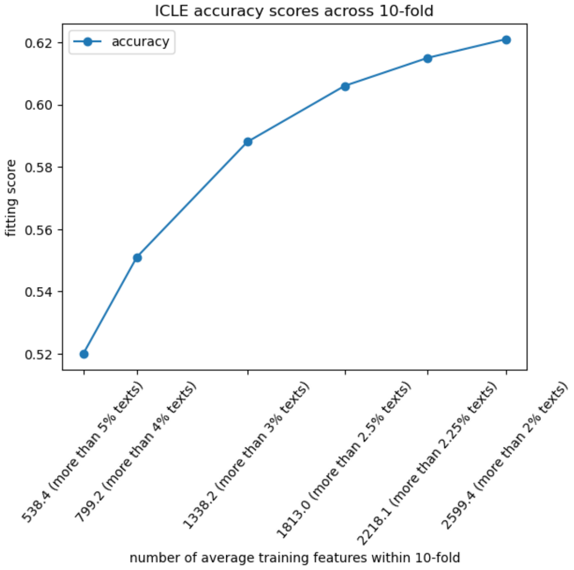


Figure 2: Model. accuracy vs. number of training features. The data is averaged across 10-fold validation.

3.3 Classification

The Ridge Classifier from *sklearn* is used in this project for three reasons. 1) The Ridge Classifier penalizes large coefficients, and such avoidance is essential for this task of lexical features, where 45% of the features in the 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. 2) It is much more time-efficient compared to other training methods that also handle sparse training data, such as support vector machine (SVM). 3) The coefficient value can reveal transfer candidates. Because the goal is to find potential collocation transfers for each L1 group, we need to identify the most characteristic features of each L1. Those with the highest coefficients are those signaling the identity of an L1 and, thus, are potential instances of collocation transfer.

4 Analysis

4.1 Model results

The fitting scores of the model demonstrate that collocations provide prediction power for NLI. Figure 2 shows the accuracy rate plotted against the number of training features. To balance between features and performance, the rest of the analysis in this paper uses about 1,800 features with an accuracy of 61%. This result outperforms baseline models using strategies of "random guessing" based on uniform probability, "most frequent label" that always selects the most frequent class, and "strati-

[1] As an example for calculation, if an L1 group contains 40 distinct prompts, and a feature occurs in five essays of prompts $prompt_1, prompt_1, prompt_1, prompt_1, prompt_2$, then the entropy value of this feature is $-\frac{4}{5} \cdot log_{40}(\frac{4}{5}) - \frac{1}{5} \cdot log_{40}(\frac{1}{5})$ = 0.136; if a feature occurs in five essays, all with the same prompt, then its entropy value is 0. A higher entropy value indicates that the feature is used in more prompts, which means that it is less likely to be influenced by topic. In this model, features with entropy values lower than 0.25 are removed.

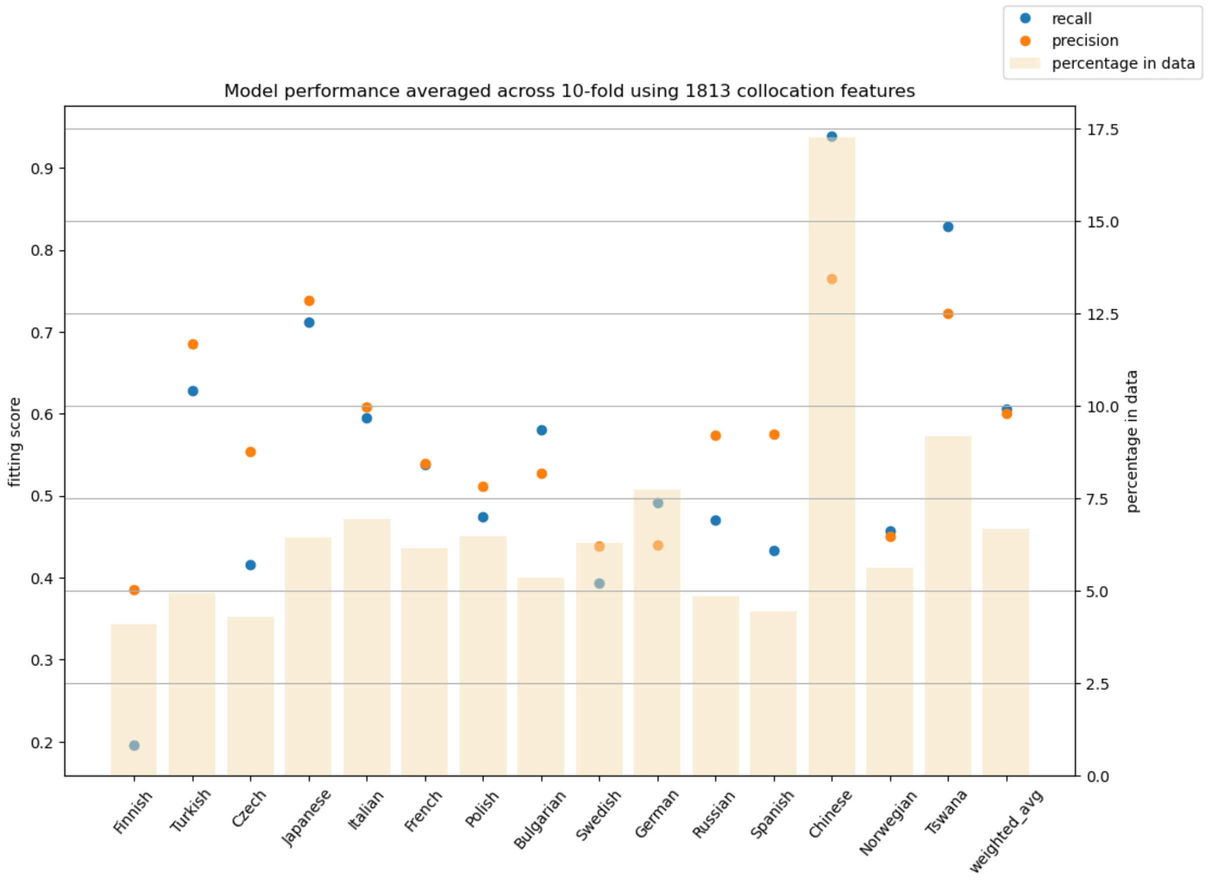


Figure 3: Model performance for each L1 averaged across 10-fold validation using 1,800 collocation features. The left x-axis represents the fitting score, and the right x-axis represents the relative sample size in percentage.

	Uniform	Most frequent	Stratified	This model (with 1,800 features)
F1	8%	5%	8%	60%
Precision	9%	3%	8%	60%
Recall	7%	17%	8%	61%
Accuracy	7%	17%	8%	61%

Table 1: Weighted average results of baseline models using strategies of uniform random guessing, most-frequent label, and "stratified", and this model with 1,800 features.

fied" (which guesses randomly based on the class distribution probability in the training data), which return accuracy rates ranging from 7% to 17%, as shown in the Table 1 [2].

A closer look at the performance of each L1 group shows that the performance varies across L1s, as shown in Figure 3. The lowest recall is Finnish (19%), and the highest is Chinese (92%). One of the reasons causing the lower fitting scores for some L1s is the unbalanced sample sizes. All L1 groups with recall rates lower than 50% (Finnish, Swedish, Norwegian, Czech, and Spanish) have below-average data sizes. Moreover, as the L1 Chinese group contributes a large portion of the data (17%), the classifier may tend to misclassify other L1 groups as L1 Chinese to achieve a better fit.

4.2 Collocation idiosyncrasies

Given the unequal performance of each L1 groups, we wonder whether the idiosyncrasies and similarities of the collocations in each group impacted the fitting result. A hierarchical clustering was performed to investigate the similarities and differences among collocations of L1 groups. For each L1, we counted the occurrences of collocates (those used by at least 2.5% of within-group samples, passing the ANOVA test, and returning an entropy value no less than 0.25), obtaining a vector documenting the frequencies of collocates from 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. The Python

[2] As our focus is on model interpretation but not model performance, we do not contrast our model with LLMs or other neural models, which may outperform our ridge classifier but are hard to interpret.

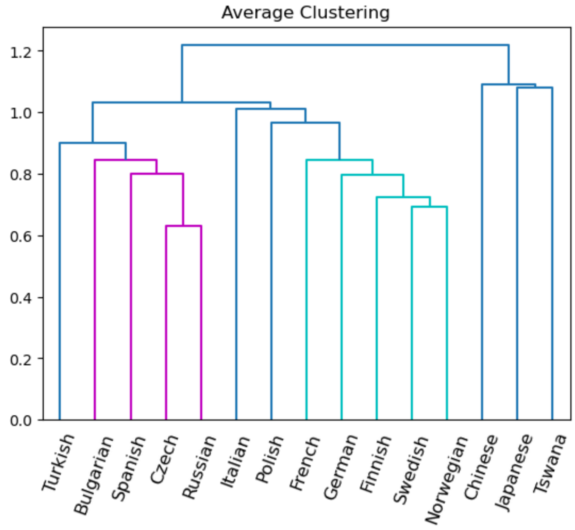


Figure 4: Hierarchical clustering dendrogram based on collocations of L1s using Ward's algorithm. Branch colors are automatically assigned by the Python package *scikit-learn*.

package *scikit-learn* (Buitinck et al., 2013) is used to implement the clustering, visualized in 4.

The clustering dendrogram, which shows the extent of similarity and difference in collocation production of these L1 groups, helps to further explain the model performance. In the dendrogram, the height of the horizontal branches where two clusters merge can be regarded as a measure of their differences, and lower height implies higher similarity. For instance, collocations produced by Norwegian, Swedish, Finnish, and German L1s is regarded as similar by the clustering method. Indeed, L1s with highly similar collocation production are relatively harder for the model to distinguish. For the German L1 group, despite a higher-than-average sample size, the classifier does not perform well (recall rate = 51%) likely because its collocations are not particularly unique, as shown by the low branch height where German is joined to other groups on the dendrogram. On the other hand, Turkish, Italian, and Japanese are joined to the dendrogram at higher branch levels, indicating a higher degree of idiosyncratic collocations these speakers produce. Unsurprisingly, the classifier performs better for these languages (recall rates 58%, 60% , and 74% , respectively), despite their medium or small sizes.

4.3 Confusion matrix

To investigate the misclassification of the model and whether this aligns with collocation similarities between groups, we plotted a normalized confusion matrix (Figure 5) that shows the percentages of predicted labels for each true label. Each row sums up to 100%. The second cell of the first row is 1.3%, which means that the classifier misclassifies 1.3% of Bulgarian writers as Chinese.

The confusion matrix aligns with the clustering dendrogram to some extent: A small-distance cluster in the middle of the dendrogram consisting of Norwegian, Swedish, Finnish, and German can explain the high misclassification rates of German as Swedish (8.2%), Swedish as German (13.5%), Finnish as German (11.3%), Finnish as Swedish (11.3%), and Finnish as Norwegian (10.4%). Another small-distance cluster, in the left part of the dendrogram, aligns with the high misclassification rates among Czech, Russian, and Bulgarian (9.1% of L1 Czech gets misclassified as Russian, and 9.5% of Russian as Bulgarian).

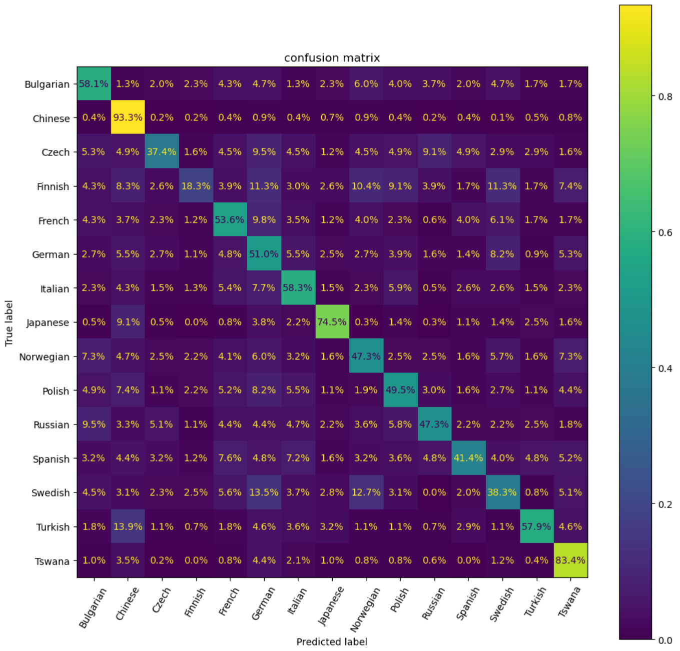
However, the clustering method is not perfect for indicating similarity distances between language groups. The adopted method, Ward's algorithm, minimizes within-cluster variance when computing the hierarchical clustering. It shows that, if Spanish is joined with the group Czech and Russian, the resulting group variance is smaller than, say, a group of Bulgarian, Czech, and Russian. However, it does not mean that Czech and Russian are the most similar groups to Spanish. In fact, Spanish L1s are most commonly misclassified as French (7.6%) and Italian (7.2%), whose similarities are not revealed in the dendrogram. This is because hierarchical clustering conveniently visualizes overall differences, but does not show the amount of differences from the perspective of each group. Future research can examine pair-wise differences in collocation production to further investigate the model misclassification and feature similarities.

4.4 Collocation features compared with previous SLA studies

The features with high coefficients are the signals the classifier identified for each L1. We compared such features with available L2 collocation studies to see if the classifier is able to find valid instances of collocation transfer. The L1s we compared to previous studies are French and Chinese, both with high classification results in this model.

4.4.1 Salient features for L1 French

We examined the top 10% features in terms of coefficient values for L1 French and compared those to the instances of collocation transfer identified



	Bulgarian	Chinese	Czech	Finnish	French	German	Italian	Japanese	Norwegian	Polish	Russian	Spanish	Swedish	Turkish	Tswana
Bulgarian	58.1%	1.3%	2.0%	2.3%	4.3%	4.7%	1.3%	2.3%	6.0%	4.0%	3.7%	2.0%	4.7%	1.7%	1.7%
Chinese	0.4%	93.3%	0.2%	0.2%	0.4%	0.9%	0.4%	0.7%	0.9%	0.4%	0.2%	0.4%	0.1%	0.5%	0.8%
Czech	5.3%	4.9%	37.4%	1.6%	4.5%	9.5%	4.5%	1.2%	4.5%	4.9%	9.1%	4.9%	2.9%	2.9%	1.6%
Finnish	4.3%	8.3%	2.6%	18.3%	3.9%	11.3%	3.0%	2.6%	10.4%	9.1%	3.9%	1.7%	11.3%	1.7%	7.4%
French	4.3%	3.7%	2.3%	1.2%	53.6%	9.8%	3.5%	1.2%	4.0%	2.3%	0.6%	4.0%	6.1%	1.7%	1.7%
German	2.7%	5.5%	2.7%	1.1%	4.8%	51.0%	5.5%	2.5%	2.7%	3.9%	1.6%	1.4%	8.2%	0.9%	5.3%
Italian	2.3%	4.3%	1.5%	1.3%	5.4%	7.7%	58.3%	1.5%	2.3%	5.9%	0.5%	2.6%	2.6%	1.5%	2.3%
Japanese	0.5%	9.1%	0.5%	0.0%	0.8%	3.8%	2.2%	74.5%	0.3%	1.4%	0.3%	1.1%	1.4%	2.5%	1.6%
Norwegian	7.3%	4.7%	2.5%	2.2%	4.1%	6.0%	3.2%	1.6%	47.3%	2.5%	2.5%	1.6%	5.7%	1.6%	7.3%
Polish	4.9%	7.4%	1.1%	2.2%	5.2%	8.2%	5.5%	1.1%	1.9%	49.5%	3.0%	1.6%	2.7%	1.1%	4.4%
Russian	9.5%	3.3%	5.1%	1.1%	4.4%	4.4%	4.7%	2.2%	3.6%	5.8%	47.3%	2.2%	2.2%	2.5%	1.8%
Spanish	3.2%	4.4%	3.2%	1.2%	7.6%	4.8%	7.2%	1.6%	3.2%	3.6%	4.8%	41.4%	4.0%	4.8%	5.2%
Swedish	4.5%	3.1%	2.3%	2.5%	5.6%	13.5%	3.7%	2.8%	12.7%	3.1%	0.0%	2.0%	38.3%	0.8%	5.1%
Turkish	1.8%	13.9%	1.1%	0.7%	1.8%	4.6%	3.6%	3.2%	1.1%	1.1%	0.7%	2.9%	1.1%	57.9%	4.6%
Tswana	1.0%	3.5%	0.2%	0.0%	0.8%	4.4%	2.1%	1.0%	0.8%	0.8%	0.6%	0.0%	1.2%	0.4%	83.4%

Figure 5: Confusion matrix of the ridge classifier with a training size of 80%. The summation of each row is 100%. Rows represent true labels, and columns represent predicted labels by the classifier.

by Paquot (2013). In Paquot's study, data from the ICLE corpus were used, and three-word bundles from L1 French writers were compared with those from 10 other L1s to see if they are used statistically differently; the frequent bundles were then triangulated with a native French corpus to validate the cause of L1 transfer. Out of the fifteen bundles identified by Paquot (2013), eight were found with high classifier coefficients in this model (i.e., deemed as characteristics of L1 French writers).

The other bundles that were identified by Paquot but did not receive high coefficients in this model were actually not included in the training features. They are likely to be excluded in the step of topic removal. While Paquot (2013) removed bundles that occurred only in the most popular topic by French writers (creation and future of Europe) but not in other topics covered by French writers, our treatment of topic influence removes more features: the use of entropy estimates the dispersion of prompts, and features that occur in more than one prompt but still only covering a small portion of all the prompts in the language group were also excluded. Therefore, the mismatch between our model results and the one by Paquot must be attributed to the different treatments of topic influence.

4.4.2 Salient features for L1 Chinese

We also investigated the intensifier-verb collocations produced by L1 Chinese to compare to a previous study by Wu and Tissari (2021). They found that Chinese learners of English produce far fewer types of intensifiers – defined as adverbs which "indicate a point on the intensity scale which may be high or low" (Quirk and Greenbaum, 1973 as cited by Wu & Tissari) compared to native English writers. As the data of the current study, the ICLE corpus, does not include native English writings, we added the LOCNESS corpus (Granger, 1998), the native counterpart compared to the ICLE corpus, to our model to identify intensifiers used by native writers. Compared to using ICLE alone, adding native data has a small impact on the fitting scores (mean f1 difference = 0.012, standard deviation of f1 difference = 0.031). Indeed, the high-coefficient features for the L1 Chinese group contain far fewer intensifiers compared to those of native writers (4 vs. 7), aligning with the findings of Wu and Tissari (2021).

Interestingly, L1 Chinese is not the only group that produces fewer types of intensifiers in their most positive adverb-verb features: among the L1 groups with the best performance in this model, L1 French and Italian groups have 6 and 5 intensifiers respectively in their high-coefficient features, while L1 Tswana and Japanese groups contain only

3. While Wu & Tissari attributed the use of limited types of intensifiers by L1 Chinese writers to the comparatively lower number of intensifiers in Chinese and the limited number of English intensifiers with direct translation equivalents in Chinese, it turns out that Tswana and Japanese writers also use fewer types of intensifiers in their collocates. In contrast, it seems that Italian and French writers have a larger repertoire of intensifiers. Potential reasons could be the comparative lack of translation-equivalent intensifiers or cognates in all Chinese, Japanese, and Tswana.

5 Discussion

This research intended to test the potential of leveraging native language identification (NLI) tasks to efficiently identify L1 transfer candidates. Focusing on collocation transfer, we show that, indeed, collocation features have predictive power to identify the L1. We asked whether the features with high positive coefficients, i.e., those deemed characteristic of each L1 group by the classifier, align with those identified in previous corpus studies. The three-word features with high coefficients for L1 French encompass those identified in a previous transfer study by Paquot (2013), except the ones excluded from our feature filtering process. The fewer types of intensifiers among the high-coefficient L1 Chinese features compared to those of native English writers confirm the findings from Wu and Tissari (2021) that Chinese writers use fewer types of intensifiers. By examining intensifiers of other well-predicted L1s (French, Italian, Tswana, and Japanese) in this model, we found a general lack of intensifier variety of non-European language L1s.

Our second research question was what caused the low fitting performance for some L1s. Using hierarchical clustering and confusion matrices, we show that, beyond the impact of small sample size, the extent of collocation idiosyncrasies affects model performance for each L1, and similarities of collocations between two L1s prompt model misclassification between them.

The current study compared features of high coefficient values to those of direct transfer patterns (French word bundles and intensifiers in Chinese). As outlined in Jarvis and Crossley (2018), by casting a wide net, NLI tasks can not only detect direct transfer patterns (i.e., those that can be found in the source language), but may also reveal indirect transfer effects, such as patterns of avoidance, or behavior that is not attested in the L1 but arises from the impact of L1 language system on L2 perception. It is interesting for future research to investigate the interpretation of indirect transfer effects based on NLI features.

6 Limitation

Since this project utilizes lexical features, which tend to occur sparsely in test data, model performance is impacted as some features of high predictive power may not be attested in the test data. The average length of essays used in this study is about 600 words, and about 45% of the features in the training data are not found in the test set. Longer texts would allow for more opportunities for each lexical feature to occur in the data, and thus are likely to improve model performance.

Although we used entropy values to mitigate the impact of topics, not all confounding factors could be removed from this study. First, the impact of the threshold of the entropy value, set at 0.25, has not been tested; It is unclear whether some collocations from topic influence survive the filtering process, especially when the information of topics is obtained only from prompts. Second, the proficiency levels in different L1 groups are not balanced in the ICLE corpus. For example, Bestgen and Granger (2011), examining argumentative essays in ICLE by L1 German, French, and Spanish, found that proficiency levels of Spanish L1s are significantly lower than that of German and French. An L1 group with low proficiency level may lead the classifier to pick out features that reflect low proficiency rather than cross-linguistic transfer (Jarvis et al., 2013).

The validity of collocation transfer also depends on the classifier's performance. For L1s with high fitting scores, such as Chinese, Japanese, and Tswana, and Italian, the confidence that their high-coefficient 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 corpus of balanced training samples and balanced proficiency levels would provide more reliable transfer candidates.

Finally, we used *SpaCy* to calculate dependency tags. However, the performance of *SpaCy* on L2 English is unknown, though its accuracy on labeled dependencies is around 90% [3].

[3] https://spacy.io/models/en#en_core_web_lg

7 Conclusion

This project demonstrates the potential of using NLI tasks to reveal collocation transfer. We find that collocations are effective features to detect L1 background, and the results provide insights into the linguistic transfer effects on collocation production. Specifically, we show that this method can capture direct collocation transfer identified by previous transfer studies, though the model performance for each L1 group is impacted by sample size and their collocation idiosyncrasies compared to other groups. While direct transfer effects can be easily confirmed by comparing features to previous transfer studies or L1 language production, the interpretation of indirect transfer effects from NLI features calls for future investigation.

References

Yves Bestgen and Sylviane Granger. 2011. Categorising spelling errors to assess l2 writing. *International journal of continuing engineering education and life-long learning*, 21(2/3):235–252.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Hakan Cangır and Philip Durrant. 2021. Cross-linguistic collocational networks in the l1 turkish–l2 english mental lexicon. *Lingua*, 258:103057.

A. Cowie. 2006. *Phraseology*, pages 579–585. Elsevier, Oxford.

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.

Dhiman Goswami, Sharanya Thilagan, Kai North, Shervin Malmasi, and Marcos Zampieri. 2024. Native language identification in texts: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3149–3160, Mexico City, Mexico. Association for Computational Linguistics.

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.

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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Radu Tudor Ionescu and Marius Popescu. 2017. Can string kernels pass the test of time in Native Language Identification? In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–234, Copenhagen, Denmark. Association for Computational Linguistics.

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

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118, Atlanta, Georgia. Association for Computational Linguistics.

Scott Jarvis and Scott A. Crossley. 2018. *The Detection-Based Approach: An Overview*, pages 1–33. Multilingual Matters, Bristol, Blue Ridge Summit.

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.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *International Conference on Intelligence and Security Informatics*, pages 209–217. Springer.

Stephen D. Krashen. 1981. *Second language acquisition and second language learning*, 1st edition. Language teaching methodology series. Pergamon Press, Oxford ;.

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 Joshua K Hartshorne. 2022. Data-driven crosslinguistic syntactic transfer in second language learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. A Deep Generative Approach to Native Language Identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1778–1783, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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, Denver, Colorado. Association for Computational Linguistics.

Ilia Markov, Vivi Nastase, and Carlo Strapparava. 2020. Exploiting native language interference for native language identification. *Natural Language Engineering*, page 1–31.

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.

J. M. Sinclair. 1991. *Collocation*. Oxford University Press, Oxford.

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.

Brent Wolter and Henrik Gyllstad. 2011. Collocational links in the l2 mental lexicon and the influence of l1 intralexical knowledge. *Applied linguistics*, 32(4):430–449.

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.