Data-driven Crosslinguistic Syntactic Transfer in Second Language Learning

Zoey Liu
Boston College
zoey.liu@bc.edu

Tiwa Eisape
Massachusetts Institute of Technology
eisape@mit.edu

Emily Prud'hommeaux
Boston College
prudhome@bc.edu

Joshua K. Hartshorne
Boston College
joshua.hartshorne@bc.edu

Abstract

Second-language (L2) learning is characterized by both positive and negative transfer from the first language (L1). However, prior psycholinguistic studies tend to focus on a few syntactic phenomena and L1-L2 pairs at a time, resulting in an incomplete picture. We apply machine learning to seven learner corpora in English and Spanish with 39 language pairs, showing that statistical models combined with simple *n*-grams of part-of-speech tags and syntactic dependency relations achieve good performance in recovering the L1, indicating structural transfer from L1 to L2. Further machine learning using a rich hand-curated linguistic feature set allows us to identify aspects of L2 linguistic structure particularly influenced by L1 (verbal morphology, average dependency tree parse depth, and headedness of clausal structures) as well as those with minimal influence (distributions of dependency relations, basic word orders, or non-projective dependencies).

Keywords: data-driven; crosslinguistic syntactic transfer; native language identification

Motivation

The lexical and structural properties shared (or not thereof) between one's native language (L1) and their second language (L2) can have notable effects on the progress of (L2) learning (Hayashi & Murphy, 2013; Ionin & Montrul, 2010; Lado, 1957; Westergaard, 2003). When L1 and L2 are typologically similar in the aspects of interest (e.g., phonological inventory, morphological structure, syntactic ordering), this potentially leads to positive transfer of linguistic knowledge from L1 to L2. For instance, children that are native speakers of Bosnian, an Indo-European (IE) language with a basic subject-verb-object word order, would have an easier time learning English, also an IE language with the same basic word order as that of Bosnian, in contrast to children whose L1 is Turkish, a Turkic language with a dominant word order of subject-object-verb (Akbarov & Đapo, 2016). Given that Spanish and English have the same inflectional morpheme to mark plurality, Spanish speakers will tend to add *-s* to nouns when learning English plurals (see also Ramirez, Chen, Geva, and Kiefer (2010)).

By contrast, linguistic knowledge of L1 can also have an inhibiting role in L2 learning, resulting in negative transfer instead. For example, Mandarin Chinese speakers, when learning English, have difficulty in using the relative clause construction (RC) correctly, since while RC is head-initial in English, it is mostly head-final in Mandarin Chinese (Chan, 2004). Learners who grow up speaking Japanese or Korean tend to pronounce the alveolar lateral approximant /l/ and the alveolar approximant /ɹ/ similarly when learning English, due to that /l/ and /ɹ/ are not phonemic in their L1.

If we take a holistic view of the literature, prior studies on linguistic transfer in second language learning broadly fall into three directions. The first one focuses on studying what individual linguistic aspects or features are transferred from L1 to L2, such as the usage of psych verbs (White et al., 1999), definite articles (Ionescu, Popescu, & Cahill, 2016) and inflectional morphology (Ramirez et al., 2010; Ramirez, Chen, Geva, & Luo, 2011). This direction mainly consists of psycholingusitic experiments, most of which usually looked at only a few linguistic features and one or two L1-L2 pairs (J. L. McDonald, 2000; O'Grady, Lee, & Kwak, 2009; Schwartz & Rovner, 2015).What's more, the linguistic aspects of interest tend to be specific to the language pairs examined; whether these features would generalize when it comes to additional L1-L2 pairs remains an open question.

The second direction attends to extensive work on Native Language Identification (NLI) in the field of natural language processing (NLP) (Berzak, Reichart, & Katz, 2014; Ionescu et al., 2016; Koppel, Schler, & Zigdon, 2005; Malmasi & Dras, 2018; Tetreault, Blanchard, & Cahill, 2013). The goal of NLI is to predict a speaker's L1 based only on the usage of their L2. In practice, NLI is commonly cast as a multi-class classification problem, where there are usually more than or around a dozen native languages to choose from. The main conjecture for NLI is that if knowledge of L1 is transferred to or has interference effects on L2, one should be able to distinguish writing, speech, or signs in L2 produced by speakers with different L1 backgrounds. In other words, L1s should be able to be predicted from characteristics of L2(s), at least to a reasonable extent. Compared to work in the first direction described above, studies along this line focused on achieving state-of-the-art accuracy rather than using NLI as a way to uncover what structural features with linguistic interpretability are transferred from L1 to L2. Therefore they tended to rely on rich feature set instead, such as character/word *n*-grams, constituent tree structures and syntactic dependency parses. In addition, most experiments studied cases where the L2 is English, with just a few notable exceptions (Portuguese (del Río Gayo, Zampieri, & Malmasi, 2018), Norwegian (Malmasi & Dras, 2018)).

The third direction makes an effort to explore what fea-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

tures are transferred from L1 to L2 in multiple language pairs, though experiments within this direction are relatively rare (Schepens, van Hout, & Jaeger, 2020). For example, with one learner corpus in English, Malmasi and Dras (2014) used NLI as lens to reveal what linguistic features have strong effects in predicting one's L1. This study found lexical n-grams to be among the most predictive features. Kulmizev et al. (2017) showed that character n-grams also have prominent roles in L1 predictions. These "bag-of-word" features, however, are quite difficult to interpret in a linguistic context. In addition, these features "give away" strong lexical cues, which are possibly specific to the language pairs, rather than being generalizable in a broader crosslinguistic context. Markov, Nastase, and Strapparava (2020) identified other useful characteristics such as punctuation, spelling errors, misuse of false cognates (e.g., using *embarrassed* as a translation for *embarazada*); yet these features are only peripherally related to linguistic structures, and they are (currently) not straightforward to characterize. In comparison, Lavrentovich (2018) examined the role of grammatical morphemes in predicting L1s for learners with different proficiency levels, but only learner corpora of English were investigated.

This study contributes to the aforementioned research gaps on linguistic transfer in second language learning. In particular, we focus on adult second language writing and ask two questions here.

(1) First, is there morphosyntactic transfer across L1-L2 pairs with different degrees of typological similarity?

(2) Second, what *interpretable* structural features are predictive of L1 *across* different language pairs?

We phrase both questions as the task of NLI. In comparison to prior work which has mainly investigated cases of which the L2 is English, we also include instances where the L2 is Spanish. Leveraging computational techniques, we explore the transfer of morphosyntactic (rather than lexical) characteristics across 7 learner corpora with 39 language pairs.

Native Language Identification

The task of NLI is hardly new in the field of NLP, with the earliest study traced back to perhaps Koppel et al. (2005). Previous work noted that being able to predict a language user's native language has practical applications in several other tasks in NLP and computational linguistics, such as authorship attribution (Wong & Dras, 2011) and grammatical error correction (Rozovskaya & Roth, 2011). Most prior studies investigating NLI have adopted the approach of combining rich feature sets with statistical classifiers such as support vector machines. Popular features include but are not limited to character/word n-grams (Kulmizev et al., 2017), constituent tree structures, syntactic dependency parses, and string kernels (Ionescu & Popescu, 2017; Ionescu et al., 2016).

Recent approaches have started to shift away from statistical classifiers to deep learning models (Devlin, Chang, Lee, & Toutanova, 2019). For instance, (Chen, 2016) performed a thorough comparisons of different neural models, including convolutional neural networks and long short-term memory networks, though these models did not appear to perform as well as statistical models. Others replaced hand-curated linguistic features with word or sentence embeddings; for example, in Steinbakken and Gambäck (2020), BERT-based models were able to achieve excellent performance for both learner corpora of English and user-generated content from Reddit (Goldin, Rabinovich, & Wintner, 2018). The current state-of-the-art system involves fine-tuning a transformer language model (GPT-2) (Lotfi, Markov, & Daelemans, 2020). These studies, again, prioritize obtaining high accuracy, with little devotion to exploring what linguistically-oriented features are prominent during NLI.

Learner corpora

In total, we used seven publicly available learner corpora, shown in Table 2. We selected data produced by language users with only one indicated L1. Data of heritage speakers was also excluded.

English learner corpora

TOEFL The ETS Corpus of Non-Native Written English (Blanchard, Tetreault, Higgins, Cahill, & Chodorow, 2014) was developed by Educational Testing Service. It consists of English essays written for the writing section of the TOEFL (Test of English as a Foreign Language) exam, produced by speakers of eleven non-English L1s; each one of them has an indicated score level (low/medium/high). After removing empty files, we acquired 12,099 essays, almost all evenly distributed across the L1s.

PELIC The University of Pittsburgh English Language Institute Corpus is a longitudinal learner corpus containing written texts as well as transcribed spoken data, all collected from different types of language classes such as grammar, listening, and reading. Based on the public repository of the corpus[1], after pre-processing, we obtained 28,314 essays produced by 1,314 students with 27 non-English L1s. The number of essays for each L1 is heavily unbalanced, ranging from 4 for Swahili to 8770 for Arabic. The students' overall language skills are divided into four proficiency levels from pre-intermediate to advanced.

WriCLE The Written Corpus of Learner English (Rollinson & Mendikoetxea, 2010) contains 711 essays in English produced by university students whose L1 is Spanish; these students have varying degrees of English proficiency.

WriCLEinf The non-academic or the informal variety of WriCLE was developed along similar efforts. The corpus includes 781 written texts also by L1 Spanish-L2 English learners; these texts fall into several different registers and genres, such as blogs, narratives, and poems.

Spanish learner corpora

CAES The Corpus de Aprendices de Espanol (Sánchez & Martínez, 2016) is comprised of 3,773 essays in Spanish writ-

[1] https://eli-data-mining-group.github.io/Pitt-ELI-Corpus/

ten by speakers of six non-English L1s. The number of essays for each L1 spans from 175 for Russian to 996 for Portuguese.

CEDEL2 The Corpus Escrito del Español L2 (Lozano, 2021) includes Spanish learner data written by students of eleven non-Spanish L1s at varying L2 proficiency levels. Pre-processing led to a total of 3034 essays; the number of essays for each L1 ranges from 58 for French to 1,921 for English.

COWS-L2H The Corpus of Written Spanish of L2 and Heritage Speakers (Davidson et al., 2020) contains written essays produced by university students who were taking lower-division Spanish classes. After pre-processing we acquired 2,129 essays from 824 students whose L1 is only English.

Experiment 1

Our first experiment seeks to address whether structural transfer from L1 to L2 exists during second language learning. In order to answer this question, we need to have morphosyntactic representations of the learner data. To that end, we relied on *n*-grams of part-of-speech (POS) tags as well as syntactic dependency relations situated within the framework of dependency grammar (Tesnière, 1959). A similar approach was applied to reconstructing linguistic phylogeny in (Berzak et al., 2014)), which showed that these features were sufficient for above-chance identification of L1; our goal for Experiment 1 is to confirm this with a much larger dataset and determine whether the effect of L1 is consistent across different L2s.

The reasons for choosing the dependency representations are twofold. First, compared to phrase-structure rules, dependency grammar is more adaptable to languages with relatively flexible word orders, which makes it easier for dependency syntax to be extended under typologically diverse contexts (see de Marneffe, Manning, Nivre, and Zeman (2021) for a recent thorough review). This is evident from the extensive efforts devoted to the development of the Universal Dependencies project (version 2.9 (Zeman et al., 2021); hereafter UD). In our cases here, we were able to find dependency treebanks for the L2 languages of interest, English, and Spanish; and these treebanks have comparable annotations. Second, the increasing availability of crosslinguistic dependency treebanks motivates the advancement of automatic dependency parsing using native monolingual data, and accordingly facilitates derivations of a variety of morphosyntactic features. Although data of this sort would not be directly comparable to L2 written essays, the parsers developed with reasonable or excellent performance for the former can be applicable to L2 writing at least within reasonable expectations (Dell'Orletta, Venturi, & Montemagni, 2011).

POS tags To obtain POS tags as well as other morphological properties such as gender and number, for each essay from the English and the Spanish learner corpora, we performed sentence segmentation, tokenization and automatic morphological annotations using Stanza (Qi, Zhang, Zhang, Bolton, & Manning, 2020), a publicly open NLP library.

Dependency parser training In order to obtain dependency parses of the essays, we trained dependency parsers for the two L2 languages, taking treebanks from UD[2]. For English, we used the UD_English-EWT treebank; for Spanish, we used UD_Spanish-AnCora. Both of these treebanks have a predefined training/development/test set.

The parser model that we adopted is Diaparser (Attardi, Sartiano, & Yu, 2021), a graph-based biaffine parser model (Dozat & Manning, 2017) which utilizes contextual embeddings and attention from transformers (Devlin et al., 2019). This model is able to directly predict dependency structures of raw texts without resorting to additional linguistic information such as POS tags or lemmas. In our experimental settings, the parser architecture was the same for the treebank of each language (Table 1): it contained three BiLSTM layers which are followed by 500 dimensional multilayer perceptron (MLP) for arc predictions and 100 dimensional MLP layer for label predictions, both with a dropout rate of 0.33. Each parser was trained using BERT embeddings (Devlin et al., 2019) with the Adam optimizer (Kingma & Ba, 2017), a batch size of 5000 and a learning rate of 2e-3. To evaluate parsing performance, we used unlabeled attachment score (UAS) and labeled attachment score (LAS) (Kübler, McDonald, & Nivre, 2009). Early stopping of training was applied based on results on the development set within the treebank. Table 1 presents parsing evaluations based on the test set; it appears that the same model architecture is able to obtain good performance for the data of the two languages. We then applied the parser of each language to its corresponding learner corpora.

Table 1: Dependency parser evaluation results on the test set of each treebank.

Language	Treebank	UAS	LAS
English	UD_English-EWT	93.02%	90.41%
Spanish	UD_Spanish-AnCora	92.73%	89.55%

Morphoyntactic features After automatically annotating the POS tags, morphological properties and syntactic dependencies for each essay, we experimented with three different morphosyntactic representations for each essay when trying to predict their native language. The first one used POS tag *n*-grams. Given an essay, we first concatenated the POS tags of all words into a sequence of POS tags. From this sequence we derived a sequence of POS tag *n*-grams ($n \leq 3$). We repeated the procedures above for all essays within a corpus, yielding a data set of POS tag *n*-gram sequences. These sequences were then transformed into numerical vectors in order, where each number is the term frequency-inverse document frequency of one *n*-gram. The second representation was similar to the first one except that it uses dependency relations instead of POS tags. The third representation simply combined the two representations above together.

Models In preliminary experiments, we explored several statistical classifiers, such as support vector machine, random

[2]Code and results in quarantine at https://github.com/zoeyliu18/crosslinguistic_nli

Table 2: Descriptive statistics of learner corpora in our experiments.

Corpus	Language	L1	*N* of documents	*N* of tokens
TOEFL	English	Arabic, Mandarin Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish	12,099	8,455,564
PELIC	English	Abrabic, Azerbaijani, Mandrain Chinese, Farsi, French, German, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Mongol, Montenegrin, Polish, Portuguese, Romanian, Russian, Spanish, Suundi, Swahili, Swedish, Taiwanese, Thai, Turkish, Vietnamese, Zulu	28,314	5,997,970
WriCLE	English	Spanish	711	1,489,386
WriCLEinf	English	Spanish	781	1,140,304
CAES	Spanish	Arabic, Mandarin Chinese, English, French, Portuguese, Russian	3,773	1,016,774
COWS-L2H	Spanish	English	2,129	1,233,116
CEDEL2	Spanish	Arabic, Mandarin Chinese, Dutch, English, French, German, Greek, Italian, Japanese, Portuguese, Russian	3,022	1,717,672

Table 3: Classification results using different morphosyntactic representations (Experiment 1) as well as hand-curated linguistic feature sets (Experiment 2).

(a) TOEFL (L2 English)

Representation	Model	Precision	Recall	$F1$
-	Majority	-	-	-
	Random	0.09	0.09	0.09
	Stratified	0.09	0.09	0.09
POS	Ridge	0.48	0.48	0.48
dependencies	Ridge	0.51	0.51	0.50
POS+dependencies	Ridge	**0.54**	**0.54**	**0.54**
feature set	Ridge	0.41	0.41	0.41

(b) PELIC (L2 English)

Representation	Model	Precision	Recall	$F1$
-	Majority	0.09	0.31	0.15
	Random	0.16	0.04	0.06
	Stratified	0.16	0.16	0.16
POS	Ridge	0.28	0.34	0.26
dependencies	Ridge	0.33	0.36	0.29
POS+dependencies	Ridge	**0.36**	**0.38**	**0.32**
feature set	Ridge	0.34	0.49	0.34

(c) CAES (L2 Spanish)

Representation	Model	Precision	Recall	$F1$
-	Majority	0.11	0.33	0.16
	Random	0.21	0.15	0.16
	Stratified	0.22	0.23	0.22
POS	Ridge	0.56	0.59	0.55
dependencies	Ridge	0.57	0.60	0.56
POS+dependencies	Ridge	**0.62**	**0.62**	**0.58**
feature set	Ridge	0.41	0.47	0.41

(d) CEDEL (L2 Spanish)

Representation	Model	Precision	Recall	$F1$
-	Majority	0.40	0.64	0.49
	Random	0.44	0.10	0.14
	Stratified	0.43	0.43	0.43
POS	Ridge	0.61	0.70	0.62
dependencies	Ridge	0.66	0.72	0.65
POS+dependencies	Ridge	0.66	0.72	0.65
feature set	Ridge	**0.72**	**0.81**	**0.75**

(e) Corpora with multiple common L1 (L2 English or Spanish)

Representation	Model	Precision	Recall	$F1$
-	Majority	0.05	0.22	0.08
	Random	0.10	0.03	0.04
	Stratified	0.11	0.10	0.11
POS	Ridge	0.41	0.43	0.38
dependencies	Ridge	0.46	0.46	0.43
POS+dependencies	Ridge	**0.49**	**0.48**	**0.45**
feature set	Ridge	0.36	0.44	0.34

(f) All seven corpora (L2 English or Spanish; all L1)

Representation	Model	Precision	Recall	$F1$
-	Majority	0.05	0.22	0.08
	Random	0.10	0.03	0.04
	Stratified	0.10	0.10	0.10
POS	Ridge	0.38	0.39	0.34
dependencies	Ridge	0.43	0.43	0.38
POS+dependencies	Ridge	**0.48**	**0.48**	**0.48**
feature set	Ridge	0.36	0.44	0.34

forest, decision trees, and the ridge classifier. Models along this line have been found to be more effective than deep learning approaches (Markov et al., 2020). We opted for the ridge classifier eventually given that it was able to achieve the best results and was more computationally efficient. We compared this classifier to three different baseline models (Table 3). For each essay, the majority baseline predicted the most frequent L1; the random baseline predicted a random L1; the stratified baseline also made random predictions but the distribution of the predictions stayed true to the actual distribution of the L1s. Note that the baseline models did not use the morphosyntactic representations described above for predictions.

Training scheme Given that we have multiple corpora for each L2, we explored three training schemes. We started with training models using data derived from each *individual corpus*. This obviously excluded corpora such as WriCLE and COWS, which only provide essays written by speakers with the same L1. The motivation for this scheme is that different corpora have different data collection procedures and research purposes. This potentially leads to domain and topic differences between corpora (Malmasi & Dras, 2018). Our goal is to see whether similar observations would hold in each corpus, meanwhile being less constrained by what data are included in the experiments.

The second training scheme combined data with a common L1 from TOEFL, PELIC, CAES and CEDEL. For the third scheme, we used all data from the seven learner corpora. All models were evaluated via 5-fold cross-validation.

Since within certain corpora the number of essays given each native language is not balanced, to be consistent across corpora, we used *weighted* precision, recall, and $F1$ as the model evaluation metrics.

Results As shown in Table 3, in spite of the specific morphosyntactic representations, overall the ridge classifiers were able to achieve good performance across every individual corpus as well as different corpus combinations. Critically, models utilizing POS tags and dependency *n*-grams together generally yielded the best performance, confirming prior findings that machine learning can detect structural transfer from L1 to L2 at scale. Of particular importance, it can do this for different L2s simultaneously, indicating that what transfers from a given L1 generalizes to different L2s.

Taking a closer look at the results, we note two patterns, both of which are within expectations from both machine learning classification perspectives and the linguistic points of view. First, not surprisingly, the model over-predicted L1s that are common within the corpus. For instance, the L1 of around 21% of L1 Chinese-L2 Spanish essays from the CAES corpus was classified as Arabic, which is the most frequent L1; looking across the Spanish L2 corpora, English was often the incorrectly predicted L1. Second – and consistent with prior work (Berzak et al., 2014) – L1s with strong typological similarities are likely to be confused; for example, when looking at all corpora, the model had difficulty distinguishing Hindi and Telugu, two predominantly head-final languages, or Spanish and Italian, two Romance languages. This further suggests that our models were picking up non-trivial linguistic manifestations of structural transfer.

Experiment 2

Experiment 1 resembles the prior work we reviewed above in that it does not tell us much about what exactly is transferred, beyond the fact that it is reflected somehow in *n*-grams of POS and dependency relations. Our second experiment aims to investigate what morphosyntactic features are potentially transferred during second language learning, using a linguistically-interpretable model. Thus in this case we opted for hand-curated features rather than just *n*-gram sequences. Since it is not currently clear what structural features would be transferred during language learning, we experimented with a wide and rich feature set, covering the characteristics of each essay at three different levels: raw texts, morphological properties, and dependency parse trees. If certain structural characteristics of the L2 writing have prominent roles in identifying the L1s, this will not only suggest transfer of knowledge for the specific characteristics, but also that this transfer holds for different language pairs (though possibly to different degrees).

Linguistic structural features Our feature set is similar to that of Brunato, Cimino, Dell'Orletta, Venturi, and Montemagni (2020) with some deviations. Given that we focused on structural characteristics, we did not include lexical features such as word *n*-grams or lexical density. For features at the raw text level, we calculated simple heuristics such as the number of sentences and words. For features at the morphological level, we relied on the morphological properties of verbs and auxiliaries used in an essay. Following the UD guidelines, we included eleven morphological features such as person, aspect, number and mood, then measured the distribution of each feature. Specifically, given each property and an individual sentence, we computed the probability of verbs with that morphological property. After repeating this step for all sentences within an essay, we derived a probabilistic distribution of this morphological feature. We then calculated the entropy (equation below) and standard deviation for the occurrences of the feature.

$$H(X) = -\sum_{i=1}^{n} P(x_i) log P(x_i)$$

With the help of POS tags, we were able to distinguish function words and content words, then calculated their respective ratios. Accordingly we measured the distributions of function and content words separately, following the procedure described above. The same was performed for auxiliary verbs as well as for lexical verbs. We then moved on to extracting syntactic features from local and global dependency trees. We included features such as the average depth of the parse tree, the degree of head-finality (i.e., the proportion of head-final dependencies within an essay) (Futrell, Levy, & Gibson, 2020), the proportion of non-projective dependencies (crossing dependencies (R. McDonald, Pereira, Ribarov, & Hajic, 2005)), and distributions of overall as well as particular dependency relations. Given the dependency annotations, we were also able to identify additional syntactic characteristics such as the main constituent order of each sentence (e.g., subject-object-verb) and the valency of the verbs, then computed the distributions of these properties.

Models Using the features described above, we trained ridge classifiers for each individual corpus as well as for when all the corpora were combined together (both using common L1s or not). Weighted precision, recall and $F1$ scores were again used as the evaluation metrics for a model's overall performance. The contribution or IMPORTANCE of each feature was computed as the difference in the model's weighted $F1$ (cross-validated) between when the full feature set is used versus when the feature was excluded. All features were then ranked based on their importance value.

Results As demonstrated in Table 3, when combined with hand-curated structural features, the results of ridge classifiers are mostly inferior to those from models using *n*-grams of POS tags and dependency relations. This is not exactly surprising based on previous findings that character or lexical *n*-grams were among the most predictive characteristics. On the other hand, the models based on the feature set still achieved much better performance compared to the three baselines. This suggests that at least some of the structural features are effective at NLI.

Looking across different corpora and training scheme, our feature ranking analysis revealed four categories of structural characteristics that are most predictive of L1s across language pairs: auxiliary verb morphology, lexical verb morphology, average dependency parse tree depth, and headedness of subordinate clauses (Table 4). For instance, when combining all corpora together, predictors with high FEATURE IMPORTANCE included the usage of past (0.02 ± 0.001) and present (0.04 ± 0.002) tense for auxiliary verbs, as well as the usage of finite (0.03 ± 0.002) and indicative (0.02 ± 0.002) verbs, suggesting that there is knowledge transfer of auxiliary verb tense and lexical verb modality from L1(s) to L2. This speaks to prior studies which have found transfer of tense-aspect morphology in English L2 learners (Ayoun & Salaberry, 2008; Muroya, 2019) as well as Spanish L2 learners (Salaberry, 2011).

Table 4: Examples of predictive and non-predictive features from hand-curated feature sets.

Predictive
distribution of past tense auxiliary verbs
distribution of present tense auxiliary verbs
distribution of finite verbs
distribution of indicative verbs

Non-predictive
distribution of specific dependency relations
distribution of basic word orders
proportion of non-projective dependency trees

In cases of dependency parse features, it appears that the average parse tree depth as well as that of subordinate clauses have notable effects in identifying L1. This finding corresponds to recent work comparing typological patterns of dependency lengths across large-scale native monolingual corpora (Futrell et al., 2020; Liu, 2020), that for instance, syntactic dependencies tend to be longer in head-final contexts. Along this line, the relative proportions of head-initial vs. head-final subordinate clauses also seem to be predictive factors. For example, given that Portuguese is more head-initial, a mixed-type language such as Mandarin Chinese (Hawkins, 1990) overall has more head-final structures. Therefore one might expect that when learning a L2, the proportion of head-final subordinate clauses is much higher in writings produced by L1 Chinese speakers (11.76%) than in texts written by L1 Portuguese speakers (7.21%), even when the L2 is predominantly head-initial; indeed, that was what we found with our experiments here.

By contrast, our results demonstrated that distributions of specific dependency relations, as well as the distribution of basic word orders and the proportion of non-projective dependencies are among the least predictive features, in the sense that including these characteristics did not lead to higher $F1$ score for the models. In term of the distribution of basic word orders, one plausible explanation is that the subject-verb-object order is porportionally much higher than other variants in both English and Spanish, with the former having more rigid word orders than the latter. Therefore it is possible that during L2 writing, the learners simply follows the basic word order, leading to a more uniform ordering distribution. Across the seven corpora, in most cases the entropy value for measuring the distribution of basic word orders is comparable and approximates 0, which in turn suggests the lack of effect for this feature.

With regards to the proportion of non-projective dependencies, given that crossing dependencies tend to lead to longer dependency lengths, which has been argued to result in processing difficulty (Gibson et al., 2019). If that were empirically true crosslinguistically, even though that some L1 in the learner corpora that we investigated, such as German and Hindi, have more non-projective dependency trees than English and Spanish, it would not be unreasonable to speculate that L2 learners possibly tend to avoid writing sentences that are too complex and cognitive-demanding, a pattern that is potentially modulated by the L2 proficiency level (see also Ouyang, Jiang, and Liu (2022) and Yan and Li (2019)).

Discussion and Conclusion

Taking data of thirty-nine language pairs from 7 learner corpora of English and Spanish, we investigated crosslingusitic syntactic transfer in adult second language writing. Using an unusually large and diverse dataset, we confirmed that the syntactic signature of L1 can be detected in L2. More critically, we demonstrate quantitatively-meaningful consistency in this structure across two L2s. Finally, we take an important first step towards measuring L1-L2 transfer on specific aspects of syntax, showing particularly large effects on verbal morphology, average dependency tree depths, and headedness of clausal structures and no evidence of effects on the distributions of particular dependency relations. Critically, unlike prior psycholinguistic studies of individual phenomena, we are able to detect these effects on a whole-language scale, addressing concerns about generalization.

These findings provide a solid foundation for theoretically-rich investigation. While our findings presented a few linguistic aspects that are potentially transferred, it remains unclear whether the transfer is positive or negative. This could be addressed by contrasting patterns in learner corpora produced by different L1 speakers to the corresponding native monolingual corpora of these L1s, in order to shed light on the nature of the structural transfer. Relatedly, we hope to add additional L2s, giving us more precision in identifying both what is consistently transferred from a particular L1 but also how the structures of the specific L1-L2 pair interact in impacting L2 learning.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grant #2127309 to the Computing Research Association for the CIFellows Project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the Computing Research Association.

References

Akbarov, A., & Đapo, L. (2016). "I am years seven old." Acquisition of English word order by Bosnian and Turkish children. *English Studies at NBU*, *2*(1), 59–73.

Attardi, G., Sartiano, D., & Yu, Z. (2021). *Diaparser attentive dependency parser.*

Ayoun, D., & Salaberry, M. R. (2008). Acquisition of English tense-aspect morphology by advanced French instructed learners. *Language Learning*, *58*(3), 555–595.

Berzak, Y., Reichart, R., & Katz, B. (2014, June). Reconstructing Native Language Typology from Foreign Language Usage. In *Proceedings of the eighteenth conference on computational natural language learning* (pp. 21–29). Ann Arbor, Michigan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W14-1603

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2014). ETS Corpus of non-native written English LDC2014t06. *Philadelphia: Linguistic Data Consortium.*

Brunato, D., Cimino, A., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2020, May). Profiling-UD: a Tool for Linguistic Profiling of Texts. In *Proceedings of the 12th language resources and evaluation conference* (pp. 7145–7151). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.883

Chan, A. Y. (2004). Syntactic transfer: Evidence from the interlanguage of Hong Kong Chinese ESL learners. *The modern language Journal*, *88*(1), 56–74.

Chen, L. (2016). *Native language identification on learner corpora*. Unpublished doctoral dissertation, M. Phil. Thesis, University of Trento, Department of Information Engineering.

Davidson, S., Yamada, A., Fernandez Mira, P., Carando, A., Sanchez Gutierrez, C. H., & Sagae, K. (2020, May). Developing NLP tools with a new corpus of learner Spanish. In *Proceedings of the 12th language resources and evaluation conference* (pp. 7238–7243). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.894

Dell'Orletta, F., Venturi, G., & Montemagni, S. (2011, June). ULISSE: an Unsupervised Algorithm for Detecting Reliable Dependency Parses. In *Proceedings of the fifteenth conference on computational natural language learning* (pp. 115–124). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W11-0314

del Río Gayo, I., Zampieri, M., & Malmasi, S. (2018, June). A Portuguese Native Language Identification Dataset. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 291–296). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-0534

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021, June). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. Retrieved from https://aclanthology.org/2021.cl-2.11

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423

Dozat, T., & Manning, C. D. (2017). Deep Biaffine Attention for Neural Dependency Parsing. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings.* OpenReview.net. Retrieved from https://openreview.net/forum?id=Hk95PK9le

Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–412.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*, *23*(5), 389–407.

Goldin, G., Rabinovich, E., & Wintner, S. (2018, October-November). Native Language Identification with User Generated Content. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3591–3601). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1395

Hawkins, J. A. (1990). A Parsing Theory of Word Order Universals. *Linguistic Inquiry*, *21*(2), 223–261.

Hayashi, Y., & Murphy, V. A. (2013). On the nature of morphological awareness in Japanese–English bilingual children: A cross-linguistic perspective. *Bilingualism: Language and Cognition*, *16*(1), 49–67.

Ionescu, R. T., & Popescu, M. (2017, September). 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* (pp. 224–234). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W17-5024

Ionescu, R. T., Popescu, M., & Cahill, A. (2016, September). String Kernels for Native Language Identification: Insights from Behind the Curtains. *Computational Linguistics*, *42*(3), 491–525. Retrieved from https://aclanthology.org/J16-3005

Ionin, T., & Montrul, S. (2010). The role of L1 transfer in the interpretation of articles with definite plurals in L2 English. *Language learning*, *60*(4), 877–925.

Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochas-*

tic Optimization.

Koppel, M., Schler, J., & Zigdon, K. (2005). Automatically determining an anonymous author's native language. In *International conference on intelligence and security informatics* (pp. 209–217).

Kübler, S., McDonald, R., & Nivre, J. (2009). Dependency parsing. *Synthesis lectures on human language technologies*, *1*(1), 1–127.

Kulmizev, A., Blankers, B., Bjerva, J., Nissim, M., van Noord, G., Plank, B., & Wieling, M. (2017, September). The Power of Character N-grams in Native Language Identification. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications* (pp. 382–389). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W17-5043`

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers.* University of Michigan Press.

Lavrentovich, A. (2018, June). Using Classifier Features to Determine Language Transfer on Morphemes. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Student research workshop* (pp. 61–66). New Orleans, Louisiana, USA: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N18-4009`

Liu, Z. (2020). Mixed evidence for crosslinguistic dependency length minimization. *STUF-Language Typology and Universals*, *73*(4), 605–633.

Lotfi, E., Markov, I., & Daelemans, W. (2020, December). A Deep Generative Approach to Native Language Identification. In *Proceedings of the 28th international conference on computational linguistics* (pp. 1778–1783). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from `https://aclanthology.org/2020.coling-main.159`

Lozano, C. (2021). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, *0*(0), 02676583211050522. Retrieved from `https://doi.org/10.1177/02676583211050522`

Malmasi, S., & Dras, M. (2014, October). Language Transfer Hypotheses with Linear SVM Weights. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1385–1390). Doha, Qatar: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D14-1144`

Malmasi, S., & Dras, M. (2018, 09). Native Language Identification With Classifier Stacking and Ensembles. *Computational Linguistics*, *44*(3), 403-446. Retrieved from `https://doi.org/10.1162/coli_a_00323`

Markov, I., Nastase, V., & Strapparava, C. (2020). Exploiting native language interference for native language identification. *Natural Language Engineering*, 1–31.

McDonald, J. L. (2000). Grammaticality judgments in a second language: Influences of age of acquisition and native language. *Applied psycholinguistics*, *21*(3), 395–423.

McDonald, R., Pereira, F., Ribarov, K., & Hajic, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 523–530).

Muroya, A. (2019). L1 Transfer in L2 Acquisition of English Verbal Morphology by Japanese Young Instructed Learners. *Languages*, *4*(1). Retrieved from `https://www.mdpi.com/2226-471X/4/1/1`

O'Grady, W., Lee, M., & Kwak, H.-Y. (2009). What the Study of Scope Can Tell us about Second Language Learning. *Journal of Pan-Pacific Association of Applied Linguistics*, *13*(1), 71–81.

Ouyang, J., Jiang, J., & Liu, H. (2022). Dependency distance measures in assessing l2 writing proficiency. *Assessing Writing*, *51*, 100603. doi: https://doi.org/10.1016/j.asw.2021.100603

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020, July). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 101–108). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.acl-demos.14`

Ramirez, G., Chen, X., Geva, E., & Kiefer, H. (2010). Morphological awareness in Spanish-speaking English language learners: Within and cross-language effects on word reading. *Reading and Writing*, *23*(3), 337–358.

Ramirez, G., Chen, X., Geva, E., & Luo, Y. (2011). Morphological awareness and word reading in English language learners: Evidence from Spanish- and Chinese-speaking children. *Applied Psycholinguistics*, *32*(3), 601–618.

Rollinson, P., & Mendikoetxea, A. (2010). Learner corpora and second language acquisiton: Introducing WriCLE. In *Analizar datos¿ describir variación:[recurso electrónico]* (p. 1).

Rozovskaya, A., & Roth, D. (2011, June). Algorithm Selection and Model Adaptation for ESL Correction Tasks. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 924–933). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P11-1093`

Salaberry, M. R. (2011). Assessing the effect of lexical aspect and grounding on the acquisition of L2 Spanish past tense morphology among L1 English speakers. *Bilingualism: Language and Cognition*, *14*(2), 184–202.

Sánchez, G. R., & Martínez, I. (2016). Learner Spanish on computer: the CAES "Corpus de Aprendices de Español" project..

Schepens, J., van Hout, R., & Jaeger, T. F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, *194*, 104056.

Schwartz, M., & Rovner, H. (2015). The acquisition of definiteness in Hebrew (L2) by bilingual preschool children (Russian-L1): A longitudinal multiple-case study. *International Journal of Bilingualism*, *19*(5), 548–571.

Steinbakken, S., & Gambäck, B. (2020, December). Native-Language Identification with Attention. In *Proceedings of the 17th international conference on natural language processing (icon)* (pp. 261–271). Indian Institute of Technology Patna, Patna, India: NLP Association of India (NLPAI). Retrieved from https://aclanthology.org/2020.icon-main.35

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck.

Tetreault, J., Blanchard, D., & Cahill, A. (2013, June). 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* (pp. 48–57). Atlanta, Georgia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W13-1706

Westergaard, M. (2003). Unlearning V2: Transfer, markedness, and the importance of input cues in the acquisition of word order in English by Norwegian children. *EUROSLA yearbook*, *3*(1), 77–101.

White, L., Brown, C., Bruhn-Garavito, J., Chen, D., Hirakawa, M., & Montrul, S. (1999). Psych verbs in second language acquisition. *The development of second language grammars: A generative approach*, *18*, 171.

Wong, S.-M. J., & Dras, M. (2011, July). Exploiting Parse Structures for Native Language Identification. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1600–1610). Edinburgh, Scotland, UK.: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D11-1148

Yan, H., & Li, Y. (2019). Beyond Length: Investigating Dependency Distance Across L2 Modalities and Proficiency Levels. *Open Linguistics*, *5*(1), 601–614.

Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., . . . Ziane, R. (2021). *Universal dependencies 2.9.* Retrieved from http://hdl.handle.net/11234/1-4611 (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University)