What shapes learner language? Exploring the roles of first language influence and developmental features with a machine learning approach

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

Stefanie Wulff
swulff@ufl.edu
University of Florida

Zoey Liu
liu.ying@ufl.edu
University of Florida

Abstract

Prior studies of Second Language Acquisition have identified two forces that shape L2 production: cross-linguistic influence and L2 developmental universals. While past studies have focused on the acquisition of target structures, the distribution of structures in learner language production may still exhibit influence from these two forces. To disentangle the two kinds of impact on the distribution of linguistic structures in natural L2 production, we took a machine learning approach, leveraging over twenty typological and syntactic complexity features from five languages (English, Spanish, Portuguese, Chinese, and Korean). By comparing linguistically informed features that characterize 1) each language based on native data, 2) each language based on L2 production from multiple L1 background, and 3) speakers of different L1 background writing in the same L2, we identified the structural features that capture the developmental universal impact and those that reflect L1 influence. Our findings reveal the complex, bidirectional, and asymmetrical interplay of language typology, cross-linguistic influence, and language development.

Keywords: learner language; cross-linguistic influence; machine learning; native language identification

Introduction

Second language (L2) learners' language production is shaped by a complex interplay of cognitive, developmental, linguistic, and social factors. Among the most extensively studied influences is the learner's first or native language (L1), commonly referred to as transfer or cross-linguistic influence (Isurin, 2005; Mitchell, Myles, & Marsden, 2019). Positive transfer occurs when L1 and L2 share similar structures, leading to facilitation (Ramirez, Chen, Geva, & Kiefer, 2010), while negative transfer results in errors when learners apply L1 rules inappropriately to L2 contexts (Odlin, 1989; Akbarov & Đapo, 2016). For instance, a Spanish speaker learning English might omit subject pronouns ("Is raining") due to pro-drop features in Spanish, or misplace adjectives ("house red") based on L1 word order. The degree of transfer depends not only on structural differences, but also on psycholinguistic factors such as proficiency, salience of features, and learner awareness (Jarvis & Pavlenko, 2008). In addition, prior work has shown that production of different L2s tend to share certain characteristics (Tessier, Sorenson Duncan, & Paradis, 2013). These characteristics shift both quantitatively and qualitatively with increasing L2 proficiency, and often bear a striking resemblance to the patterns of children acquiring their first language. Rather than being derived (solely) from the L1 or learner input, these systematic production patterns may emerge as learners construct internal hypotheses about grammatical rules of the L2. For example, overgeneralization (e.g., "comed" for "came" in English L2 production) parallels patterns seen in L1 acquisition and suggests an underlying learning mechanism sensitive to L1 or input regularities. Similarly, developmental sequences—such as the progressive acquisition of English morphemes (Brown, 1973; Dulay & Burt, 1974) or negation patterns—tend to appear in consistent orders across learners regardless of background. These phenomena highlight that learner language is a dynamic, evolving system governed by both developmental linguistic universals as well as influence from the L1s.

This raises the question: can we *distinguish the two alternative sources of impact on L2 production*? In other words, can we distinguish L2 production patterns that arise as developmental universals during L2 learning in general, versus patterns that are possibly attributed to L1 transfer effects instead? While prior work on developmental universals and L1 transfer shares similar focus on whether target structures have been acquired by learners (defined as an emergence of the target structures in the learner production), we here try to disentangle the two sources by looking at the distribution of multiple structural features, which, even when"acquired" by learners, may display different distributions due to L1 influence or impact from L2 developmental universals. We take inspiration from Jarvis and Crossley (2012), who advocate for a machine learning approach. Provided the right (amounts of) data, they argue, machine classifiers can open the scope of investigation to a wide array of features, allow for multiple group comparisons at once, and help detect elusive and indirect cross-linguistic effects. Elaborating on this, we perform L1 and L2 language classification separately and compare the classifications to identify developmental universal features.

Overview of Experiments

To address our research questions, we need native and L2 production data across languages, and the resulting dataset(s) should be sufficient for data-driven, machine-learning classification experiments. While data obtained from psycholinguistic paradigms offer insights for targeted phenomena through controlled studies and possibly probe into individual learning differences (McDonald, 2000; O'Grady, Lee, & Kwak, 2009; Ramirez et al., 2010; White et al., 1999),

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

these studies can be limited in terms of the number of linguistic phenomena, stimuli, participants studied, which leads to comparatively smaller datasets for larger-scale quantitative examinations. To address these potential limitations, we resort to *written* corpora as the test case. This is not to say that written data takes priority over other types of production data, but rather that it is more suitable for our current study.

Additionally, given our focus on linguistic features and their roles in L2 learning from a cross-linguistic perspective, we need to rely on linguistic characteristics that apply to different languages. Since we are studying written corpora, we selected syntactic features based on the framework of dependency grammar (Tesnière, 1959; de Marneffe, Manning, Nivre, & Zeman, 2021), for three reasons. First, the dependency grammar framework has been widely adopted for typological studies (Baylor, Ploeger, & Bjerva, 2024; Berzak, Reichart, & Katz, 2014), and previous experiments have shown that certain dependency-based features that are well-motivated by psycho-linguistic theories (e.g., the Dependency Locality Theory (Gibson et al., 2000)) can capture cross-linguistic structural tendencies (Futrell, Levy, & Gibson, 2020). Second, the Universal Dependencies project (UD) (Nivre et al., 2020) contains over 100 dependency treebanks with native data and gold-standard annotations for a variety of languages. These annotations were conducted following the UD guidelines, which offer cross-linguistic comparability for annotation standards. Lastly, the existence of UD has motivated continuous development of off-the-shelf natural language processing tools such as lemmatizers, part-of-speech (POS) taggers, and dependency parsers. Since almost no L2 corpora have pre-existing manual annotations (see Data Preprocessing and Features), these tools help with our automated preprocessing of the L2 data, which renders the resulting annotations at least somewhat comparable to those of the native data, and lets us extract the same set of morphosyntactic features from all corpora. For these reasons, we adopted dependency syntax and the UD-style annotations.

We performed three classification tasks using distribution information of a wide set of sentence-level structural features.

- **L → L classification (Experiment 1):** A classifier is trained to identify a native speaker's language based on structural features attested in natural production data. Features that turn out to be predictive point to structural differences among languages.
- **L2 → L2 classification (Experiment 2):** A classifier is trained to identify an L2 speaker's language based on their L2 language production using the same structural features as in the first model. Features that turn out to be predictive point to structural differences among these languages showcased through general L2 production. Further, we can compare this classifier to the first classifier to find out differences between L2 and native production (i.e., developmental universals).
- **L2 → L1 classification (Experiment 3):** This classifier is trained to identify a speaker's L1 background based on the same structural features as attested in L2 production data. Predictive features in this classifier likely point to cross-linguistic influence or transfer from the speaker's L1.

Data Preprocessing and Features

For L2 written corpora, we searched through the 29 publicly accessible datasets of learner essays from Liu, Qin, Yang, and Hartshorne (2025). As part of our goals is to investigate cross-linguistic influences, preferably, we would like an L1 to appear in multiple L2 corpora with a good amount of samples, and the native speakers of such L2 languages to conversely contribute to the L2 corpora of this L1 language (e.g., samples of L1 English in L2 Spanish and L2 Portuguese, and samples of L1 Portuguese and L1 Spanish in L2 English.)

Based on available data, we experimented with UD (considered as native data) and L2 data of five different languages: English, Spanish, Portuguese, Chinese, and Korean. For the native data of every language, we adopted correspondingly one treebank from UD version 2 (Nivre et al., 2020), of which the metadata clearly states the genres of the data. We chose treebanks consisting of the genres news, blogs and Wikipedia texts for consistency. The L2 data came from a variety of corpora (see Table 1). Notice that UD treebanks provide random sentences extracted from written texts, thus we consider a sample from UD treebanks as a sentence; on the other hand, in the L2 corpora, each sample is a piece of essay.

For L2 data, we used the open-source library STANZA (Qi, Zhang, Zhang, Bolton, & Manning, 2020) to automatically obtain POS tags and syntactic dependency relations. For each essay, we only included sentences that have more than five tokens and contain tokens other than punctuations besides the root; additionally, essays with fewer than 30 tokens were excluded. No such processing was needed for native data because they already come with gold-standard annotations.

Structural Features We employed twenty-six features that have been used to characterize language typology (e.g., dependency length, tree height, head direction, determiner types, SVO order) (Levshina, 2019; Futrell et al., 2020) as well as other syntactic complexity features (e.g., noun phrase complexity and subclause length) identified in previous (L2) research (Jiang, Bi, & Liu, 2019; Ortega, 2015):

- **Sentence dependency length:** Human language production tends to minimize dependency length (Gibson et al., 2019); shorter dependency length potentially facilitates online comprehension and processing efficiency (Fedorenko, Woodbury, & Gibson, 2013). Here this feature is measured as the overall dependency length of a sentence as well as that of a noun phrase (NP).
- **Sentence tree height:** After converting each sentence to a hierarchical dependency tree, this feature corresponds to the number of levels between the root of the sentence and the children at the lowest syntactic level. The motivation for this feature (Zhang & Liu, 2018) is similar to that of dependency length.

Table 1: Corpora used in this study.

	Corpora used in L→L classification	L2 Corpora
English	EWT - A Gold Standard UD Corpus for English (Bies et al., 2012)	TOEFL - The ETS Corpus of Non-Native Written English (Blanchard, Tetreault, Higgins, Cahill, & Chodorow, 2014) PELIC - The University of Pittsburgh English Language Institute Corpus (Juffs, Han, & Naismith, 2020) WriCLE - The Written Corpus of Learner English (Rollinson & Mendikoetxea, 2010) WriCLEinf - the non-academic or informal counterpart of WriCLE CLC - The Cambridge Learner Corpus (Yannakoudakis, Briscoe, & Medlock, 2011) ICNALE - The International Corpus Network of Asian Learners of English (Ishikawa, 2013) ICLE - The International Corpus of Learner English (Granger, Dagneaux, Meunier, Paquot, et al., 2009) BAWE - The British Academic Written English Corpus (Nesi, Gardner, Thompson, & Wickens, 2008) ArabCC - The Arab Academic College of Education MOECS - The Corpus of Multilingual Opinion Essays by College Students
Spanish	GSD - The Google Spanish Universal Dependency treebank (Nivre et al., 2020)	CAES - The Corpus de Aprendices de Espa¯nol (Miaschi et al., 2020) CEDEL2 - The Corpus Escrito del Espa¯nol (Lozano, 2021) COWS-L2H - The Corpus of Written Spanish of L2 and Heritage Speakers (Davidson et al., 2020)
Portuguese	GSD - The Google Portuguese Universal Dependency Treebank (Nivre et al., 2020)	COPLE2, PEAPLE & Leiria - The Portuguese Native language Identification Dataset (del Río Gayo, Zampieri, & Malmasi, 2018)
Chinese	GSD - The Google Traditional Chinese Universal Dependency Treebank (Nivre et al., 2020)	TOCFL - The Test of Chinese as a Foreign Language
Korean	GSD - The Google Korean Universal Dependency Treebank (Chun, Han, Hwang, & Choi, 2018)	KLC - The Korean Learner Corpus (Lee, Tseng, & Chang, 2018)

- **Length and frequencies of individual types of subclauses:** L1 typology may affect the interpretation of L2 subclauses when such L2 structures are difficult for learners (e.g., Kanno (2007) found that beginner L2 Japanese learners, when not able to parse Japanese relative clauses, turn to L1 word-order as a parsing strategy).
- **Dependency length, length, and frequency of noun phrases:** Phrasal complexity is a hallmark of advanced proficiency and professional/academic language (Biber, Gray, & Poonpon, 2011). We include NP complexity to capture any stylistic L1 influence.
- **Head directionality in sentences and NPs:** Futrell et al. (2020) proposed that head-finality (the proportion of head-final dependencies in a phrase or sentence) describes language typology on a gradient continuum; they showed further the interaction between head-finality and dependency length such that languages that are more head-final tend to have longer dependencies (Liu, 2020). Here we adopted the same measure at both the sentence and NP-levels.
- **Word orders of subject, object, and verb:** Constituent word orders have long been used to classify language typology (Dryer, 1997). Perceived word order similarity is likely to prompt transfer from the L1 (Rothman, 2010). Here we focus on the ordering of subject (S), object (O), and verb (V) at both the main constituent level (where the verb is the root of the sentence) as well as at the clausal level. For each sentence, we measured the proportion of SV/VS, VO/OV, and SO/OS.

Experiment 1

To ensure that our structural features are effective at capturing the typology of the languages included in our study, i.e., distinguishing these languages, we built an L→ L classification with the native data only. The process is as follows: (1) for each of the five gold-standard treebanks taken from the UD project directly for our five languages of interest, we first extracted the aforementioned feature values from every sentence in the treebank; (2) we combined all feature values from the five treebanks as input to our classifier, which was trained to predict the corresponding language of each sentence.

If the classifier does a good job, this means the structural features can capture typological differences between languages. We chose the multinomial logistic Lasso regression as it automatically performs feature selection (Tibshirani, 1996). To remove any unwanted influence of uneven treebank sizes on the classification results (treebank sizes here ranged from 15K sentences (Mandarin GSD) to 48K sentences (Spanish GSD)), the samples (i.e., sentences, in the case of UD treebanks) for each language were bootstrapped (randomly sampled with replacement) to achieve equal sample sizes: First, we took the sample size of the language with largest sample size and multiplied it by a positive constant larger or equal to 1 (in our case, 2); then, for each language, samples were randomly selected with replacement until the selected ones reached the size obtained in the first step.

For all experiments, we compared our feature-based classifiers with one baseline: *uniform* baseline, which randomly assigns a language (out of the five) to a sentence. Two other baselines are commonly used: *stratified* baseline, which predicts language labels based on distribution, and *most frequent* baseline, which always predicts the most frequent label. However, because our samples are balanced for each label, the *stratified* baseline return the same results as the *uniform* baseline, and the *most frequent* baseline is meaningless (there is no "most frequent" label). Weighted average precision, recall, and F1 scores were used to measure model performance, though our analysis focused on F1 scores.

Results The weighted average F1 value was 0.77 (baseline F1 = 0.20), indicating that the model performed well in identifying the languages based on the structural features we chose. See Table 2, second column, for the model performance.

The model results demonstrate the impact of language typology. Portuguese and Spanish tend to be misclassified as each other (16% and 20% respectively), and they can be both misclassified as English to a lesser degree (about 10%). Similarly, the most common incorrect labels assigned to English are Portuguese (17%) and Spanish (13%).

Table 2: Fitting scores of L→L and L2→L2 classifications.

	L-L			L2-L2		
	Recall	Precision	F1	Recall	Precision	F1
English	0.61	0.68	0.64	0.94	0.93	0.94
Spanish	0.68	0.67	0.67	0.69	0.72	0.71
Portuguese	0.71	0.66	0.68	0.75	0.73	0.74
Korean	1	0.99	1	1	1	1
Mandarin Chinese	0.86	0.84	0.85	0.99	0.99	0.99
Overall Weighted Avg	***0.77***	***0.77***	***0.77***	***0.88***	***0.88***	***0.88***
Baseline (uniform)	0.20	0.20	0.20	0.20	0.20	0.20

A closer look at the distinctive features reveals the most and least useful features for this classification task. Useful

	Positive features		Negative features	
	L->L	L2 -> L2	L->L	L2 -> L2
Chinese	proportion of head-final structures in sentence	proportion of head-final structures in sentence	proportion of S/V	number of clausal subject
	sentence dependency length	number of subclauses	noun phrase dependency length	number of noun phrase
	number of subclauses	number of clausal complement	proportion of S/O	proportion of S/V
	content/function word ratio	sentence maximum dependency length	number of open clausal complement	proportion of S/O
	proportion of V/O	length of noun phrase	sentence tree height	sentence length
	number of clausal complement	sentence tree height	length of adverbial clause	content/function word ratio
English	proportion of S/V	proportion of head-final structures in noun phrase	proportion of S/O	number of clausal complement
	proportion of V/O	number of adnominal clause	number of clausal subject	sentence tree height
	proportion of head-final structures in noun phrase	proportion of V/O	number of noun phrase	proportion of S/O
	noun phrase dependency length	number of noun phrase	content/function word ratio	content/function word ratio
	number of open clausal complement	length of adverbial clause	sentence length	proportion of S/V
	length of open clausal complement	sentence length	length of noun phrase	noun phrase dependency length
Korean	content/function word ratio	number of subclauses	proportion of V/O	proportion of V/O
	sentence dependency length	sentence tree height	length of open clausal complement	sentence length
	sentence maximum dependency length	sentence maximum dependency length	noun phrase dependency length	length of subclause
	proportion of S/O	number of clausal complement	sentence tree height	length of open clausal complement
	number of subclauses	content/function word ratio		length of adverbial clause
	number of noun phrase	length of adnominal clause		
Portuguese	proportion of S/V	number of clausal subject	proportion of head-final structures in noun phrase	proportion of head-final structures in sentence
	sentence tree height	noun phrase dependency length	number of noun phrase	proportion of head-final structures in noun phrase
	proportion of S/O	proportion of S/V	proportion of head-final structures in sentence	number of adverbial clause
	length of adverbial clause	proportion of V/O	number of adverbial clause	sentence dependency length
	number of open clausal complement	number of open clausal complement	number of clausal complement	number of clausal complement
		length of adverbial clause	number of subclauses	number of noun phrase
Spanish	noun phrase dependency length	number of clausal subject	proportion of head-final structures in noun phrase	proportion of head-final structures in sentence
	sentence tree height	proportion of S/O	number of clausal complement	proportion of head-final structures in noun phrase
	number of clausal subject	noun phrase dependency length	proportion of head-final structures in sentence	number of open clausal complement
	length of clausal subject	content/function word ratio	proportion of V/O	proportion of V/O
		number of noun phrase	number of adnominal clause	sentence dependency length
			number of subclauses	number of subclauses

Figure 1: Distinctive features in L→L and L2→L2 classifications, sorted by coefficient values. Features different in the two classifications are shaded yellow, features shared but with opposing coefficient signs are shaded blue.

here is defined as "being distinctive to tell apart a language from others". The more useful features, or distinctive features for three or four of the five languages investigated here, are word order features, number of subclauses (individual types and overall number), sentence tree height, dependency length, proportion of head final structures, and content function word ratios. Length features are among the least useful: some are only distinctive for one of five languages (clausal subject length for Spanish, sentence length for English, and noun phrase length for Korean), and the other length features are not distinctive for any (word length of adjective clauses, clausal complements, all subclauses). The feature of proportion of definite vs. indefinite articles is not useful either.

Experiment 2

We next asked whether L2 production of a language, regardless of L1 background, converges to some extent; or to put differently, whether there are common production patterns emerging in a language produced by learners. To answer this question, we performed L2 → L2 classification, where a model was trained to identify an L2 language based on the structural feature values attested in production. If the model can outperform the baselines at identifying the L2 languages based on L2 data, this means that L2 data of a particular language share structural similarities, and can thus be identified by the model. The procedure was identical to Experiment 1, except that the data were L2 corpora instead of UD treebanks.

As we wanted to look at L2 production in general, we were

keen to avoid any undue influence from a particular L1 background. To achieve this, we balanced the contributions of the various L1 backgrounds to each L2 sample using a bootstrap method (randomly sampling data without replacement) so that the data reflect the "average" look of a piece of L2 production across learners with varied L1 backgrounds. Any L1/L2 with less than 50 samples (the number of essays) was excluded. The number of L1s per L2 ranged from 32 (English) to 9 (Chinese) (Table 3). As in Experiment 1, the overall sample size of each L2 language was balanced to avoid potential sample size effects on classification results.

Table 3: L1s of each L2 used in L2→L2 classification. The sample size of each L1 in each L2 is balanced using bootstrapping. L1/L2s with >50 samples (essays) are excluded.

L2	L1s
English	Chinese (Cantonese, Mandarin), Norwegian, Hungarian, Bulgarian, Polish, Turkish, Macedonian, Finnish, Japanese, Greek, Tswana, Italian, Portuguese, Persian, French, Lithuanian, Dutch, Serbian, Korean, Swedish, German, Spanish, Punjabi, Urdu, Russian, Czech, Arabic, Hindi, Telugu, Thai, Catalan, Indonesian
Spanish	Mandarin Chinese, English, Arabic, Portuguese, Russian, French, Japanese, Greek, Italian, German, Dutch
Portuguese	French, Mandarin Chinese, German, Romanian, English, Russian, Italian, Spanish, Polish, Japanese, Korean
Korean	Vietnamese, English, French, Chinese (Mandarin, Cantonese), Japanese, Khmer, Mongolian, Spanish, Russian, Indonesian, Portuguese, Kazakh, Swedish, Nepali, Tagalog, Uzbek, Sinhalese, Italian, Burmese, German, Kyrgyz, Bengali, Arabic, Turkish,
Chinese	Japanese, Vietnamese, Indonesian, Korean, English, Thai, French, Spanish, German

Results The weighted average F1 was 0.88 (baseline F1 = 0.2; Table 2). This high result means that L2 production of a particular language, captured by our structural features and regardless of L1 background, does indeed converge.

Similar to Experiment 1, the model here also shows influence of language typology. As expected, the model confuses Spanish and Portuguese, the two most typologically similar languages in our data. However, unlike the L→L classification, the L2→L2 model performs much better with English (F1 of 0.9 vs. 0.6). A potential reason is that the EWT English corpus used in L→L classification may contain data from speakers growing up with influences from various language backgrounds, given the popularity of English as the lingua franca in many fields. Consequently, the UD English corpus may in fact be less homogeneous than desired.

The least and most useful features that tell apart the L2 languages were similar to those in the L→L classification. The least useful features were word length of phrases and clauses (e.g., clausal subject, clausal complements, noun phrases, adjective clauses) and proportion of definite vs. indefinite determiners; the most useful ones were word order features, number of subclauses (individual types and overall number), sentence tree height, dependency length, proportion of head final structures, and content function word ratios.

L→L vs. L2→L2 Distinctive Features

Even though the most and least distinctive features in L→L and L2→L2 classifications are almost identical, the set of distinctive features for each individual language shows that L1 and L2 productions are far from identical. We compared the distinctive features in the L→L and L2→L2 classification tasks by comparing the rankings of their coefficients from the respective classifiers (since the coefficient values themselves are not directly comparable). For all five languages, the distinctive features for the two classifications are partially different. For Korean, both L→L and L2→L2 classifications show traits of higher ranking for content function ratio, max dependency length of sentences, and number of overall subclauses. L2 Korean differs from L1 Korean: when compared with other L2 languages, L2 Korean stands out with higher sentence tree height, higher number of clausal complements, and a higher content function ratio, which are not characterizing L1 Korean when compared to other L1 productions. For Chinese, the mean content function ratio feature is assigned a positive sign in L→L classification, but a negative sign in L2→L2 classification. Similar observations hold for all five languages (see Figure 1). In summary, while L1 and L2 production share certain features, they differ with regard to others. We interpret these differences to arise from the L2 learning process, not from L1 influence, because the L2 data included a wide range of L1s (9-32 L1s depending on the L2s).

Experiment 3

Experiment 2 identified likely developmental universals, as they characterize a language produced in the L2 independent of speakers' L1 background. To identify those features that likely arise from L1 influence, we performed L2→L1 classification, which is also known as native language identification (NLI (Goswami, Thilagan, North, Malmasi, & Zampieri, 2024)). For each L2 (English, Spanish, Portuguese, Korean, Chinese), we built a multinomial Lasso logistic regression to predict the speakers' L1 based on the structural feature values derived from the L2 data; the approach is similar to prior experiments. We here included a smaller set of L1s (the top four to six L1s, depending on their sample size) than in Experiment 2, so that each L1-L2 combination has a large number of samples to adequately represent the average structural choices made by learners of this L1 background. As before, we used bootstrapping to balance sample sizes.

Results The model performance for each L2 is presented in Table 4. As expected, for all L1/L2s, the F1 scores of our classifier are above those of the baseline, indicating that the features capture cross-linguistic influence.

Language typology interfered with model performance: typologically related L1s tend to be confused more. For instance, in L2 Portuguese, over 40% of L1 Japanese gets misclassified as Korean, and over 30% of L1 Chinese gets misclassified as Japanese, yielding comparatively worse model performance for these two L1s. Similar results were found for L1 English, L1 Portuguese, and L1 Spanish.

Unlike the comparison made between L→L and L2→L2 classification that showcases features that arise from L2 learning in general, the distinctive features in this L2→L1 classification reflect L1 influence. For each classifier, the data is L2 production from different L1s in the same L2; thus, the classifier should not pick up on general L2 learning features, but only L1 influence features (Jarvis & Crossley, 2012).

The informativeness of features in L2→L1 classification differs from our previous two experiments. While content function word ratios, number of individual types of subclauses and the overall number of subclauses, and proportion of head final structures continue to be useful, some features that are not useful in Experiment 1 and 2 (length of clauses, sentences, and NPs) are distinctive for a number of L1s and L2s. Conversely, unlike in the previous two experiments, word order features are less useful here. This is expected, because now a classifier examines production data in the same language (just different L1 backgrounds), and due to the grammatical constraints of a particular language, speakers may enjoy less freedom in choosing the word order of subject, verb, object, compared to the other structural features. The proportion of definite vs. indefinite determiners is distinctive only for a few L1/L2s, once again one of the less useful features across our experiments.

One interesting observation is that there are only two positive distinctive features for L1 Spanish-L2 Portuguese, meaning that not many features can tell L1 Spanish apart from other L1s. This could be due to the typological closeness of these two languages, causing the L2 Portuguese production to bear less L1 (Spanish) influence. Interestingly, however, this is not true the other way around in L1 Portuguese-L2 Spanish. For now, we interpret this finding as an example of the often-observed asymmetries in cross-linguistic influence between language pairs, due to asymmetries in markedness, learners' perceived language distance, or language dominance, among other factors (Kellerman, 1979; Romano, 2021).

L2	L1	Recall	Precision	F1
English	Japanese	0.47	0.41	0.44
	Korean	0.32	0.38	0.35
	Chinese	0.29	0.34	0.31
	Portuguese	0.48	0.41	0.44
	Spanish	0.42	0.44	0.43
	Arabic	0.48	0.46	0.47
	Weighted avg	***0.41***	***0.41***	***0.41***
	Baseline (uniform)	0.17	0.17	0.17
Spanish	Japanese	0.67	0.59	0.63
	Chinese	0.28	0.38	0.32
	Portuguese	0.40	0.41	0.41
	English	0.58	0.50	0.53
	Arabic	0.48	0.49	0.48
	Weighted avg	***0.48***	***0.47***	***0.47***
	Baseline (uniform)	0.20	0.20	0.20
Portuguese	Japanese	0.34	0.30	0.32
	Chinese	0.21	0.25	0.23
	Korean	0.45	0.46	0.45
	English	0.25	0.32	0.28
	Spanish	0.46	0.42	0.43
	Weighted avg	***0.34***	***0.35***	***0.34***
	Baseline (uniform)	0.20	0.20	0.20
Korean	Japanese	0.49	0.40	0.44
	Chinese	0.19	0.31	0.23
	Arabic	0.42	0.37	0.39
	English	0.14	0.24	0.17
	Spanish	0.33	0.35	0.34
	Portuguese	0.56	0.39	0.46
	Weighted avg	***0.36***	***0.34***	***0.34***
	Baseline (uniform)	0.16	0.16	0.16
Mandarin Chinese	Japanese	0.33	0.33	0.33
	Korean	0.28	0.36	0.32
	English	0.34	0.34	0.34
	Spanish	0.49	0.39	0.43
	Weighted avg	***0.36***	***0.35***	***0.35***
	Baseline (uniform)	0.25	0.25	0.25

Table 4: Classification result for NLI.

Discussion

We investigated the question *how to disentangle two alternative sources impacting L2 production: developmental universals and L1 influence*. We first performed a language classification task (L→L) to show that our selected structural features are able to distinguish languages (English, Spanish, Portuguese, Korean, and Mandarin Chinese). Next, we conducted an L2→L2 classification that distinguished different L2s based on L2 production using the same feature set as before. This experiment confirmed that learner language is a distinct variety, as different L2s can be told apart in spite of the fact that data of each L2 comprised contributions by speakers of varied L1 backgrounds. By comparing the features distinctive in the L2→L2 and L→L classification, we found that L2 production generally looks different from native production; we attributed these differences to the effects of L2 learning itself in general rather than L1 influence, which aligns with previous studies on developmental universals (Meisel, Clahsen, & Pienemann, 1981).

In the third experiment, we carried out L2→L1 classification, where an L1 was identified based on L2 production. Here, distinctive features likely show L1 influence instead of the impact from general L2 learning, considering that the essays for each classifier are all in the same L2 and the main difference among them is their L1 background. We found that some features (length of clauses, sentences, and NPs) that are not informative in the other two classifications turned out useful when predicting L1s based on L2s. This means that, though these features may not characterize the L1s, they emerge as L1 influence on L2 production.

The partial overlap in distinctive features between Experiments 1-and-2, 2-and-3, and 1-and-3 reflects the complex, bidirectional, and asymmetrical interplay of language typology, cross-linguistic influence, and language development. One future direction is a deeper analysis of language-specific patterns of cross-linguistic influence and developmental universals to understand why, for a L2 language, its developmental universal features and cross-linguistic features emerge the way they are, and if a prediction of whether these features would emerge during L2 learning based on language structures would be possible.

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.

Baylor, E., Ploeger, E., & Bjerva, J. (2024, March). Multilingual gradient word-order typology from Universal Dependencies. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 2: Short papers)* (pp. 42–49). St. Julian's, Malta: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.eacl-short.6/ doi: 10.18653/v1/2024.eacl-short.6

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

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in l2 writing development? *TESOL Quarterly*, *45*(1), 5-35. doi: https://doi.org/10.5054/tq.2011.244483

Bies, A., et al. (2012). *English Web Treebank LDC2012t13*. Philadelphia: Linguistic Data Consortium. (Web Download)

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2014). ETS Corpus of non-native written English LDC2014T06. *Philadelphia: Linguistic Data Consortium*.

Brown, R. (1973). *A first language: the early stages*. Harvard University Press.

Chun, J., Han, N.-R., Hwang, J. D., & Choi, J. D. (2018). *Building universal dependency treebanks in korean.* Miyazaki, Japan.

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

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.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021, June). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308.

Dryer, M. S. (1997). On the six-way word order typology. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, *21*(1), 69–103.

Dulay, H., & Burt, M. (1974). Natural sequences in child second language acquisition. *Language Learning*, *24*, 37–53.

Fedorenko, E., Woodbury, R., & Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive Science*, *37*(2), 378–394.

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.

Gibson, E., et al. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, *2000*, 95–126.

Goswami, D., Thilagan, S., North, K., Malmasi, S., & Zampieri, M. (2024, June). Native language identification in texts: A survey. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 3149–3160). Mexico City, Mexico: Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.173

Granger, S., Dagneaux, E., Meunier, F., Paquot, M., et al. (2009). *International corpus of learner English* (Vol. 2). Presses universitaires de Louvain Louvain-la-Neuve.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*, *1*(1), 91–118.

Isurin, L. (2005). Cross linguistic transfer in word order: Evidence from L1 forgetting and L2 acquisition. In *Proceedings of the 4th international symposium on bilingualism* (Vol. 1115, p. 1130). Somerville, MA: Cascadilla Press.

Jarvis, S., & Crossley, S. A. (Eds.). (2012). *Approaching language transfer through text classification: Explorations in the detection-based approach*. Multilingual Matters Channel View Publications.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*.

Jiang, J., Bi, P., & Liu, H. (2019). Syntactic complexity development in the writings of efl learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing*, *46*, 100666.

Kanno, K. (2007). Factors affecting the processing of japanese relative clauses by l2 learners. *Studies in Second Language Acquisition*, *29*(2), 197–218. Retrieved 2026-01-19, from http://www.jstor.org/stable/44487148

Kellerman, E. (1979). Transfer and non-transfer: Where we are now. *Studies in Second Language Acquisition*, *2*, 37-57.

Lee, L.-H., Tseng, Y.-H., & Chang, L.-P. (2018, May).

Building a TOCFL learner corpus for Chinese grammatical error diagnosis. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1363

Levshina, N. (2019). Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, *23*(3), 533–572.

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

Liu, Z., Qin, W., Yang, H., & Hartshorne, J. K. (2025). Studying cross-linguistic structural transfer in second language learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 47, pp. 4061–4070).

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

Meisel, J. M., Clahsen, H., & Pienemann, M. (1981). On determining developmental stages in natural second language acquisition. *Studies in Second Language Acquisition*, *3*(2), 109–135. doi: 10.1017/S0272263100004137

Miaschi, A., Davidson, S., Brunato, D., Dell'Orletta, F., Sagae, K., Sanchez-Gutierrez, C. H., & Venturi, G. (2020, July). Tracking the evolution of written language competence in L2 Spanish learners. In *Proceedings of the fifteenth workshop on innovative use of nlp for building educational applications* (pp. 92–101). Seattle, WA, USA â†' Online: Association for Computational Linguistics.

Mitchell, R., Myles, F., & Marsden, E. (2019). *Second language learning theories*. Routledge.

Nesi, H., Gardner, S., Thompson, P., & Wickens, P. (2008). *British academic written English corpus.* (Oxford Text Archive)

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., ... Zeman, D. (2020, May). Universal Dependencies v2: An evergrowing multilingual treebank collection. In N. Calzolari et al. (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 4034–4043). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.497/

Odlin, T. (1989). Language transfer: Cross-linguistic influence in language learning. *Cambridge University Press*.

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.

Ortega, L. (2015). Syntactic complexity in l2 writing: Progress and expansion. *Journal of second language writing*, *29*, 82–94.

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.

Romano, F. (2021). L1 versus dominant language transfer effects in l2 and heritage speakers of italian: A structural priming study. *Applied Linguistics*, *42*, 945–969.

Rothman, J. (2010). On the typological economy of syntactic transfer: Word order and relative clause high/low attachment preference in l3 brazilian portuguese. *IRAL - International Review of Applied Linguistics in Language Teaching*.

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

Tessier, A.-M., Sorenson Duncan, T., & Paradis, J. (2013). Developmental trends and L1 effects in early L2 learners' onset cluster production. *Bilingualism: Language and Cognition*, *16*(3), 663–681. doi: 10.1017/S136672891200048X

Tibshirani, R. (1996, 1). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288. doi: 10.1111/j.2517-6161.1996.tb02080.x

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.

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011, June). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 180–189). Portland, Oregon, USA: Association for Computational Linguistics.

Zhang, H., & Liu, H. (2018). Interrelations among dependency tree widths, heights and sentence lengths. *Quantitative analysis of dependency structures*, *72*, 31–52.