Studying Cross-linguistic Structural Transfer in Second Language Learning

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

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

Wenshuo Qin
wqin6@jh.edu
Johns Hopkins University

Joshua K. Hartshorne
joshua.hartshorne@hey.com
MGH Institute of Health Professions

Abstract

Adults who learn a new language often report feeling that their first language gets in the way. Systematic effects of the first language on additional languages would have straightforward implications for both theory and pedagogical practice *if* they could be adequately characterized. Unfortunately, this has been shown to be challenging. Languages are vast and complex, and there are a very large number of them. Thus, most studies focus on a few narrowly-defined phenomena and one or two language pairs. The potential complexity of the phenomenon and the sparsity of the observations conspire to make it difficult to establish clear patterns. We present whole-language analyses of the morphosyntax of 133,659 second-language essays spanning 273 L1-L2 pairs. We find clear, consistent effects of the L1 on the morphosyntax of the L2, independent of the L2. We find that not all aspects of morphosyntax are equally informative about the L1, suggesting avenues for more precisely specifying how and why L1 influences L2.

Keywords: L2 learning; morphosyntactic transfer; machine learning; native language identification

Introduction

Many learners of additional languages report feeling that the morphosyntax of their first language (L1) has a systematic effect on learning the morphosyntax of a second language (L2). (Note we use 'L2' loosely, including third languages, etc.) There are a number of theoretical accounts as to why this might be the case (Unsworth, 2010; Epstein, Flynn, & Martohardjono, 1996; Jarvis, 2017; Mitchell, Myles, & Marsden, 2019). Indeed, a number of researchers have suggested that L2 knowledge is parasitic on L1 knowledge. An influential early account suggested that second-language learners could not reset parameters of Universal Grammar learned for their L1 (DuPlessis, Solin, Travis, & White, 1987). Another posits that the L1 constrains the hypothesis space that learners consider for the L2 (Bley-Vroman, 1990). Working in a very different tradition, Hernandez, Li, and MacWhinney (2005) suggest that late-learners of an L2 (as opposed to simultaneous bilinguals) are unable to successfully segregate the representations for the different languages, and that the L2's representations end up being parasitic on the L1's. Other researchers are more optimistic, arguing that while early learners initially base the L2 on the L1, they can reanalyze the input in a way that ultimately does not depend on the L1 (B. D. Schwartz & Sprouse, 1996). Yet other researchers have noted that even if the L1 and L2 representations are entirely distinct, the L1 may still interfere with the L2 as speakers attempt to select the correct language to use (Ahn & Ferreira, 2024).

This brief review hardly covers the space of theoretical accounts, but the point is that systematic effects of the L1 on the morphosyntax of the L2 would have straightforward implications for both scientific theory and pedagogical practice *if* we could identify what that systematic effect was. This has been demonstrated to be challenging. First, most studies only focus on a small number of narrowly-defined phenomena in a handful of languages at best (Mitchell et al., 2019; White et al., 1999; Ramirez, Chen, Geva, & Luo, 2011; A. I. Schwartz, Kroll, & Diaz, 2007; J. L. McDonald, 2000; Lee, Tseng, & Chang, 2018; Ionescu, Popescu, & Cahill, 2016; Hartsuiker & Bernolet, 2017), making it difficult to establish patterns or make well-justified generalizations. A handful of studies have compared speakers from a variety of L1 backgrounds learning English as an L2 (Yun, Li, Li, & Hartshorne, 2023; Malmasi & Dras, 2014; Berzak, Reichart, & Katz, 2014; Malmasi & Dras, 2015; Schepens, van Hout, & Jaeger, 2020). However, they have been focused on detecting L1-L2 influence, with limited or no characterization of the influence. In any case, the focus on L2 English limits ability to generalize, since effects may be idiosyncratic to the specifics of English morphosyntax or the unique sociolinguistic status of English.

Liu, Eisape, Prud'hommeaux, and Hartshorne (2022) take an important step towards addressing the aforementioned limitations, pioneering a method to make a nuanced quantitative assessment of L1-L2 transfer across many syntactic phenomena simultaneously and for many language pairs. Like Malmasi and Dras (2014) and Berzak et al. (2014), they use machine learning to identify morphosyntactic features in L2 essays that can identify the writer's L1. While most work along these lines is focused on the practical challenge of identifying the writer's L1 ["native language identification"; (Koppel, Schler, & Zigdon, 2005)], the features of L2 morphosyntax that are characteristic of specific L1s provides a quantiative window into the effect of L1 on L2. However, machine learning representations can be difficult to interpret, so Liu et al. (2022) used a more interpretable ridge regression. Critically, they simultaneously analyzed the effects of L1s on two different L2s (English and Spanish); thus, their results highlight systematic effects of L1 on L2 that are common across those L2s. They report certain features of verbal morphology are the most predictive of learners' L1s, whereas characteristics such as main word order and distributions of dependency relations are among the least predictive.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

L2	Corpora
English	**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; Granger, 2003) **BAWE** - The British Academic Written English Corpus (Nesi, Gardner, Thompson, & Wickens, 2008) **Gachon** - The Gachon Learner Corpus (Carlstrom & Price, 2012-2014) **ArabCC** - The Arab Academic College of Education **MOECS** - The Corpus of Multilingual Opinion Essays by College Students
German	**MERLIN_German** - The German section of the MERLIN corpus
Norwegian	**ASK** - The Language Learner Corpus of Norwegian as a Second Language (Tenfjord, Meurer, & Hofland, 2006)
Icelandic	**IceL2EC** - The Icelandic L2 Error Corpus (Ingason, Stefánsdóttir, Arnardóttir, Xu, & Glišić, 2021)
Spanish	**CAES** - The Corpus de Aprendices de Español (Miaschi et al., 2020) **CEDEL2** - The Corpus Escrito del Español (Lozano, 2021) **COWS-L2H** - The Corpus of Written Spanish of L2 and Heritage Speakers (Davidson et al., 2020)
Portuguese	**COPLE2**, **PEAPLE** & **Leiria** - The Portuguese Native language Identification Dataset (del Río Gayo, Zampieri, & Malmasi, 2018)
Italian	**UD_Italian-Valico** - dataset from (Di Nuovo, Bosco, Mazzei, & Sanguinetti, 2019) **MERLIN_Italian** - The Italian section of the MERLIN corpus.
Czech	**Czesl** - The Learner Corpus of Czech (Hana, Rosen, Škodová, & Štindlová, 2010) **MERLIN_Czech** - The Czech section of the MERLIN corpus (Wisniewski et al., 2018)
Croatian	**CroLTec** - The Croatian Learner Text Corpus (Preradović, Berać, & Boras, 2015)
Latvian	**LaVA** - The Latvian Language Learner Corpus (Darģis, Auziņa, & Levāne-Petrova, 2018)
Finnish	**LAS2** - The Corpus of Advanced Learner Finnish LAS2 (Ivaska, 2014)
Chinese	**TOCFL** - The Test of Chinese as a Foreign Language (Lee et al., 2018)
Korean	**KLC** - The Korean Learner Corpus

Table 1: Learner corpora in our experiments.

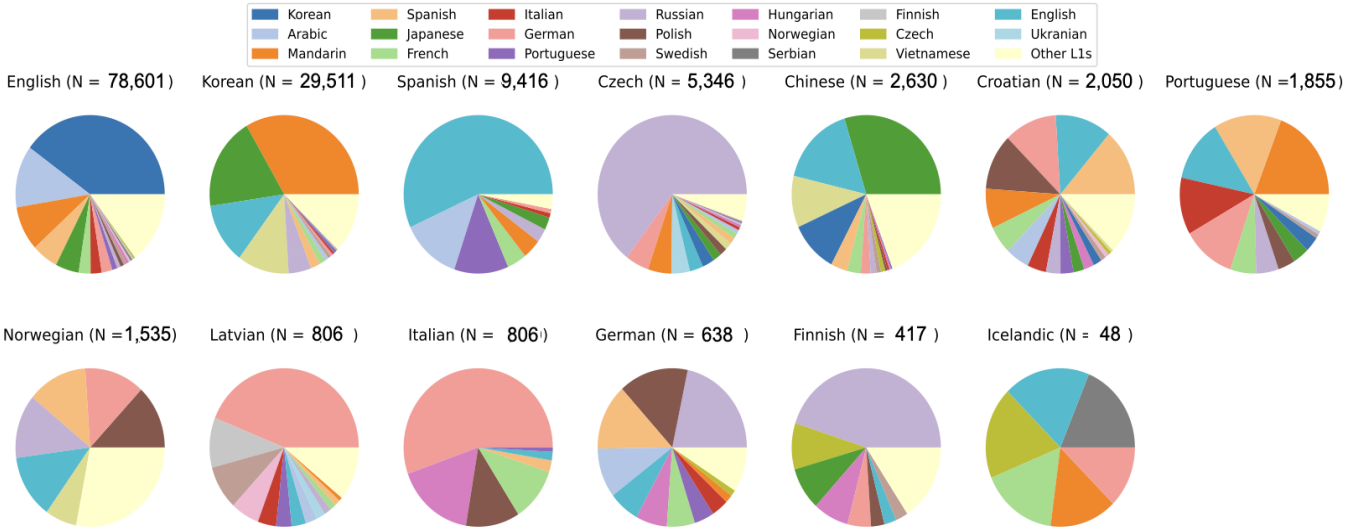

Figure 1: Visualizations of the L1 distributions for each L2 in our experiments; the value of *N* refers to the total number of essays collected from written learner corpora for a given L2.

Liu et al. (2022) was limited, however, in that only 39 L1-L2 pairs were considered, only 2 closely-related L2s were studied, and only 9 L1s were in common across the L2s. In the present study, we expand to 273 L1-L2 pairs with a total of 13 L2s from 4 language families (Table 1; Figure 1). This puts us in a much better position to assess whether the morphosyntax of the L1 has consistent effects on L2 learning independent of the L2. While results cannot be conclusive — we are still sampling only a fraction of L1s and L2s — the empirical scope of this study far exceeds all prior work. Moreover, the dataset is much larger, even for L2 English (78,601 vs. 41,905 essays) and mildly for L2 Spanish (9,416 vs. 8,924 essays). Thus, statistical findings should be fairly robust, at least for the languages sampled.

Possible outcomes include: 1) the L1 cannot be reliably identified from the morphosyntax of the L2, in which case there are no consistent L1-L2 transfer effects that generalize across the languages studied here; 2) the L1 can be reliably identified, and all morphosyntactic features investigated are useful for identifying it; or 3) the L1 can be reliably identified, but only based on certain morphosyntactic features, such as those implicated in productive rules (Morgan-Short

& Ullman, 2022) or in Universal Grammar-related phenomena (DuPlessis et al., 1987). (Note that we are not taking a stance on whether there are productive rules or a Universal Grammar, though of course the results here could inform such debates.)

Overview of Experiments

We compiled a large database of essays written in L2s. We used automatic methods to extract part-of-speech tags and syntactic dependency relations. In Experiment 1, we use standard machine learning classifiers to confirm that we can identify the L1s, both when training on each L2 separately and when training on all simultaneously. However, this method does not provide much insight into whether only some aspects of morphosyntax are helpful in classifying the L1. Following Liu et al. (2022), in Experiment 2 we derive a set of handcrafted syntactic features to use for training and investigate which are most predictive of L1.

Limitations: There are several limitations to this work that are important to bear in mind. First, we are analyzing natural behavior, not representation. That is, we can tell whether a learner of a specific L2 produces native-like morphosyntax but not necessarily whether they do so in the same way a bilingual might. For that, one would need to use deliberate syntactic priming, neural measures, or other psycholinguistic methods that provide a more direct assay of underlying linguistic representations (Ahn & Ferreira, 2024; Hartsuiker & Bernolet, 2017; Salamoura & Williams, 2007; Bermúdez-Margaretto et al., 2022). Relatedly, the theories briefly reviewed above generally do not make clear predictions about natural behavior. That would require computational models that do not yet exist and which would likely require further specification of the theories first. Thus, our study is largely data-driven rather than theory-driven, and connecting to theory will require more work. Our hope is to inspire such work (see Discussion).

Second, the syntactic features used in Experiment 2 by definition must be present in all L2s. Thus, this approach is not intended to detect, for instance, the effect of L1 on acquisition of language-specific structures, because not all the L2s necessarily have those structures. Even there, our choice of features is a function of what we could think of and straightforwardly compute, which is likely less than everything that could be computed. Thus, our findings are necessarily about a (potentially proper) subset of L1-L2 influences on L2 morphosyntax. While we can imagine methods that are not subject to this limitation, they all require more sophisticated models of language learning than currently exist or are even possible, given the current level of specificity of theories and the available computational techniques. Again, we hope the present study will help spur such work.

Third, we are analyzing written essays, a format that allows the participant more time for reflection and revision than speech. Depending on the L2, performance may be influenced by the writing system itself and, conversely, only minimally by issues of phonology. Thus, we likely miss some aspects of L1-L2 influence and pick up on issues specific to literacy. Relevantly, studying free behavior means that participants have the flexibility to avoid constructions they find confusing or otherwise simply fail to produce utterances that would be particularly informative. This is a fundamental tradeoff between studying natural behavior (the thing science actually wants to explain) and controlled laboratory experiments targeted to test specific phenomena (and are thus particularly informative). Because most of prior research has been the latter, we feel the conducting the former is currently high-value, but it certainly is not a replacement.

Fourth, there are any number of practical concerns. While it is likely that this paper covers more L1-L2 pairs than all previous work combined, it is still a tiny fraction of the world's languages. Similarly, we cannot rule out the possibility that L1 is confounded with essay content or style (perhaps speakers of particular L1s just like certain topics more), which is then reflected in syntactic patterns. We are currently developing methods to address this concern, but it is not trivial. Likewise, our statistical analyses are, like any analyses, imperfect. In particular, results may be affected by multicollinearity, the amount of variability of each feature, the performance of the parser (particularly with respect to parsing non-normative syntax), and so on. Again, our hope here is merely to be less wrong and push things forwards.

Learner Corpora and Preprocessing

We obtained 29 learner written corpora covering 13 L2s, spanning the Indo-European ($N = 10$), Uralic ($N = 1$), Sino-Tibetan ($N = 1$), and Koreanic ($N = 1$) language families (Table 1; Figure 1). From each corpus we selected data produced by learners with a single specified L1; data of heritage speakers was not included.

We analyzed morphosyntax using dependency grammar (Tesnière, 1959; Osborne, 2014). While there are many competing frameworks with arguments in their favor (Langacker, 1997; Wundt, 1910; Chomsky, 1956; Steedman, 2014; Pollard & Sag, 1994), dependency grammar has the decided advantage of being well-defined for a wide variety of languages (Zeman et al., 2024), including all 13 of the L2s in our study. This is in part because dependency grammar is robust to languages with flexible word orders (de Marneffe, Manning, Nivre, & Zeman, 2021). We used the open-source library STANZA (Qi, Zhang, Zhang, Bolton, & Manning, 2020) to automatically derive part-of-speech (POS) tags, morphological feature annotations such as tense, number, gender, etc., and dependency relations.

From each processed essay, we selected sentences with at least five tokens. We further excluded sentences with only punctuations besides the root token. After these filtering steps, we excluded any essay with fewer than 30 tokens, and also removed L1-L2 language pairs with fewer than five essays. Preprocessing led to 133,659 essays.

While demographics about the writers are minimal – limiting some kinds of analyses – the dataset is rich in L1s and L2s. We present visualizations of the distribution of the L1s for each of the 13 L2s in our data in Figure 1. The total number of essays for each L2 ranges from 78,601 for English and 29,511 for Korean, to 417 for Finnish and 48 for Icelandic. Of the 273 unique L1-L2 pairs, the most frequent are Korean and English (N=31,389), Arabic and English (N=10,225), and Mandarin and Korean (N=9,826). In comparison, the distributions of most other L1-L2 pairs are more sparse, with 135 (49.45%) pairs having fewer than 50 essays.

Experiment 1

With the automatically annotated learner data, we first examine whether systematic morphosyntactic transfer can be detected across L1-L2 language pairs.

Deriving morphosyntactic representations For each sentence in a given essay, we extracted POS tag trigrams. For instance, in the following example sentence

```
I like cheese .
PRON VERB NOUN PUNCT
nsubj root obj punct
```

the POS tag trigrams are *PRON+VERB+NOUN* and *VERB+NOUN+PUNCT*. We do the same for dependency relations (*nsubj+root+obj, root+obj+punct*). We concatenate the POS and dependency trigrams for all sentences in the essay in a linear order, resulting in a structural representation of the essay.

The simplicity of these representations means helps ensure that our findings are data-driven and reasonably theory-neutral. Moreover, prior research has obtained good native language identification results with similar (if slightly richer) representations (Berzak et al., 2014; Berzak, Nakamura, Flynn, & Katz, 2017).

L1 classification In initial experimentation, we employed different models, including statistical classifiers such as ridge classifier, and neural networks such as convolutional neural network. We chose ridge classifier eventually because of its efficient computation, as well as that it yields performance comparable to, or even better than, that of the neural networks. This is perhaps not surprising given the data-hungry nature of the latter (Markov, Nastase, & Strapparava, 2020). We compared the performance of the ridge classifier to three baselines: the *majority* baseline, which predicts the most frequent L1 in the learner data; the *random* baseline, which predicts L1s randomly; and the *stratified* baseline, which predicts L1s based on their original distributions in the corpora.

We trained and tested classifiers on each L2 separately, as well as on the entire dataset simultaneously. While each learner corpus presumably has its own data collection process, along with different essay topics, writing instructions, and numbers of writers, it is not possible to know the relevant information for sure since some corpora lack detailed documentation and metadata. To investigate whether the observations from the two aforementioned classification schemes hold across different writing settings, we also built classifiers for each individual learner corpus, excluding corpora with only one L1. All classifiers were evaluated with 3-fold cross-validation; we used weighted $F1$ score as a measure of classifier performance.

Results

As shown in Table 2, for each of the 13 L2s, trigram sequences of POS tags and dependency relations, coupled with ridge classifiers, are able to predict L1s with good performance, outperforming all three baselines. This pattern holds for each individual L2 (Table 2) and also on each individual corpus (not shown due to space limitation). While performance on the full set of L2s is sometimes lower than for models trained for specific L2s, this is in part because the classification problem is harder (there are more L1s in the full dataset). In any case, performance is still quite good.

These numbers collectively provide support that structural transfer exists consistently across L1-L2 language pairs, a finding that perhaps is made even stronger by the fact that our morphosyntactic representations here are quite simple. Additionally, these results also contribute to existing literature on native language identification (see Goswami, Thilagan, North, Malmasi, and Zampieri (2024) for a review) which commonly adopts machine learning with distributional linguistic information to predict L1.

	Exp. 1 (Trigrams)				Exp. 2 (Features)
L2	**Majority**	**Random**	**Stratified**	***Model***	***Model***
English	0.23	0.03	0.19	**0.48**	0.30
German	0.08	0.07	0.13	**0.24**	0.21
Norwegian	0.02	0.11	0.11	**0.25**	0.21
Icelandic	0.02	0.14	0.17	**0.57**	0.45
Spanish	0.41	0.11	0.36	**0.65**	0.53
Portuguese	0.06	0.08	0.11	**0.30**	0.20
Italian	0.39	0.19	0.37	**0.62**	0.54
Czech	0.50	0.04	0.42	**0.53**	0.51
Croatian	0.03	0.04	0.09	**0.23**	0.16
Latvian	0.27	0.08	0.22	**0.34**	0.31
Finnish	0.28	0.09	0.23	**0.38**	0.34
Chinese	0.14	0.06	0.16	**0.31**	0.21
Korean	0.17	0.04	0.18	**0.35**	0.25
all	0.09	0.02	0.11	**0.42**	0.26

Table 2: L1 classification $F1$ scores (out of 1) for Exp. 1 (trigrams) and Exp. 2 (derived morphosyntactic features). For Exp. 1, we also show the results for three baselines: Majority, Random, Stratified. Results are shown for models trained and tested on individual L2s as well as a model trained and tested on all L2s (final row).

Experiment 2

Having demonstrated that morphosyntactic transfer can be detected reliably across L1-L2 pairs, now we turn to our second research question, which is whether L1 influences specific aspects of L2 morphosyntax more than others. While trigram sequences of POS tags and dependency relations are sufficient to show that structural transfer effects exist and that such effects are generalizable across L1-L2 pairs, they are not easy to interpret.

To address this question, we opt for a different approach via designing a rich hand-curated feature set. Our feature set

is largely similar to that of Liu et al. (2022) with some modifications (see also Brunato, Cimino, Dell'Orletta, Venturi, and Montemagni (2020)). The feature set combines structure information at the textual, morphological, and syntactic levels.

Textual features: Features at the raw-text level were mostly heuristic; examples included the numbers of sentences and words, average sentence length, the number of unique POS tags and dependency relations, etc. Since we are interested in structural transfer, we purposefully excluded lexical features such as type-token ratio (Richards, 1987) or lexical density (Malvern, Richards, Chipere, & Durán, 2004), which can be indicative of language proficiency and development.

Morphological features: For features at the morphological level, we included the morphological features of function words and content words (separating all tokens into these two categories only in this case based on their POS tags). We also studied the morphology of lexical verbs and auxiliaries, such as mood, number, tense, and aspect. In addition, we also examined the morphological properties of adjective (degree of comparison, e.g., comparative, superlative), determiners (definiteness), nouns (singularity) and numbers (cardinality), as well as pronouns (e.g., case, number, person). These features were automatically derived from Stanza (Qi et al., 2020) for each L2; all the annotations followed the UD standards, hence comparable across languages.

Given each morphological feature, we measured its probabilistic distribution using entropy (Eq (1)), an information-theoretic (Cover & Thomas, 2006) measure that has been widely used as a proxy of linguistic complexity (Futrell, 2024; Juola, 2008) and has also been applied in L2 learning contexts (Sun & Wang, 2021).

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

Others have shown entropy and its mathematical derivations can be indicative of online processing behaviors (Linzen & Jaeger, 2014; Pimentel, Meister, Wilcox, Levy, & Cotterell, 2023) and typological tendencies (Ackerman & Malouf, 2013; Levshina, 2019). Here we take entropy as a metric to reflect how variable the usage of a feature is; a higher value of entropy corresponds to more variation. Our conjecture is that there should be distinguishable differences (though possibly to different degrees when facing different L1-L2 pairs) in the amount of variation for certain features in L2 production that can help with identification of L1s. Aside from entropy and other distributional measures such as standard deviation, we also computed the production ratio of each feature.

Syntactic features At the syntactic level, we extracted features from the local and global dependency parse trees of the essays (Malmasi & Dras, 2014). Selected features included the distributions of both individual and overall dependency relations; to approximate these distributions, we also used entropy. Additionally, we included features that can characterize some information of language typology by previous literature, such as the distribution of main word orders that involve subject, head verb, and object; the depth of the dependency parse tree; average dependency length (Liu, 2020); the proportion of non-projective dependencies (a sentence having crossing dependency arcs) (R. McDonald, Pereira, Ribarov, & Hajic, 2005) which are more frequently found in languages with more flexible word orders (Dyer, 2017) and are considered to be linguistically more complex as these structures tend to incur online processing difficulty (Husain & Vasishth, 2015; Gibson et al., 2019); and the proportion of head-final dependencies (where the syntactic dependent appears before its head) (Futrell, Levy, & Gibson, 2020).

L1 classification: We used the feature set described above for L1 classification, again with the ridge classifier. Again, we trained an omnibus model as well as separate models for each L2. Classifier performance was indexed by weighted $F1$ score.

To identify which features are most informative about L1, we used the measure PERMUTATION IMPORTANCE from the Python package `scikit-learn`, which computes the difference in model performance when all the values for a given feature are randomly shuffled. A larger absolute value of permutation importance indicates that a certain feature is more "important" and accordingly, more predictive of L1, whereas a smaller value corresponds to the opposite.

While our initial hand-curated feature set contains a total of 240 structural features, in practice, the feature set downsizes to different extents for every L2 (e.g., 80 features for Latvian; 28 for Korean) since the values of some of the features turn out to be zeros due to either non-existence of certain morphological properties in the data or data sparsity (e.g., case markers of verb mood in the Czech learner corpora). Taking that into account, we expect the feature set to not yield models with high performance. But, if at least some of the features are potentially transferred during L2 learning and have notable effects on predicting L1, we expect the feature set to at least result in classification performance better than the baselines from Experiment 1.

Results

Results from Table 2 align with our expectations: $F1$ scores from our hand-curated feature sets outperform all the baselines from Experiment 1 for all L2s.

Predictive features
the aspect of verbs
the form of verb (e.g., finite, infinite)
the person of auxiliary
the proportion of verb usage
Non-predictive features
the main constituent order
the number of auxiliary
the head directionality of subordinate clause

Table 3: The most predictive vs. non-predictive features from the hand-curated feature set.

Of all the features investigated, the most predictive ones fall into three categories at different levels. At the raw text

level, the average sentence length seems to be effective. At the morphological level, the aspect and form of verbs and the person of auxiliaries appear to be the most predictive; this aligns (partially) with prior literature (Montrul, 2011) which suggests the role of inflectional morphology in both L2 and heritage language learning. At the syntactic level, the average dependency length as well as the average depth of the dependency parse trees of an essay are among the most indicative of L1. This is not surprising given findings from recent work on syntactic typology (Futrell et al., 2020; Liu, 2020) that there are systematic differences in the overall dependency lengths between different language types.

In comparison, the main constituent order, the number feature of auxiliary, as well as the proportion of head-final dependencies in subordinate clauses (Table 3) are among the factors that are the least predictive.

Discussion

In a large dataset of learner essays involving 273 L1-L2 pairs, we see consistent signatures of L1 on L2, such that the L1 could be identified based on morphosyntactic patterns in the L2, independent of L2. While not shocking, it is also not necessarily *a priori* true; it could have been possible that transfer effects are an idiosyncratic interaction of L1 and L2, such that there would be no consistent "grammatical accent". Intriguingly, we see evidence that systematic grammatical accents are specific to specific aspects of morphosyntax.

Note that our method does not clearly distinguish between negative transfer and compensation. That is, L1 influence on the morphosyntax of an L2 could be due to the learner incorrectly using the L1 grammar (negative transfer). Alternatively, it could be compensation: the learner has not (yet) acquired the relevant morphosyntax and so compensates by translating from the L1. The difference is important for understanding critical period effects: the widely-observed but still poorly-understood fact that older learners rarely acquire a new language to the same level of proficiency as native speakers (J. K. Hartshorne, 2022; J. Hartshorne, 2024; Johnson & Newport, 1991; Chen & Hartshorne, 2021).

We outlined some of the more salient limitations of our work above. One that should not be understated is the potential that non-predictive features are non-predictive due to collinearity or insufficient variation in the dataset. Nor would we discount the possibility of confounds, with speakers of different L1s writing about different topics, either due to cultural differences or imbalances across the 29 corpora. While we have no particular reason to believe this would affect morphosyntax, we also have no good arguments that it cannot. These problems of analysis are not trivial; what we present here is the best we have managed so far.

A practical limitation not mentioned above is interpreting the patterns in the results for the 240 features. For instance, it is perhaps notable that all the most predictive features involve verbs, but then so does one of the least-predictive (Table 3). Ideally, we would partition the features based on some grammatical theory and ask whether categorically different aspects of morphosyntax (as classified by that theory) are differentially affected by L1-L2 transfer. However, theories are not typically specified at that level. The lack of an effect on main constituent order or head directionality of subordinate clauses would seem to militate against accounts in which language learning involves setting parameters of Universal Grammar that cannot then be reset for an L2, since typical proposals include parameters affecting both those phenomena.[1] But even for such theories, it can be difficult to determine for some arbitrary syntactic phenomenon, whether it counts as part of Universal Grammar parameters, particularly since lists of proposed Universal Grammar parameters are understood to be incomplete. Moreover, language is a complex dynamical system with myriad opportunities for compensation: working out what the effects of difficulties with, say, learning head directionality would actually be not trivial. This becomes even more difficult for other formalisms, which are often even less specific.

This consideration highlights the need for more precise theories and models of language acquisition that can be applied at the level of an entire language, rather than individual phenomena in isolation or for small artificial languages. We believe this is increasingly within reach. For instance, Constantinescu, Pimentel, Cotterell, and Warstadt (2025) recently investigated L2 acquisition in large language models, finding clear critical period effects — essentially an updated version of Hernandez et al. (2005). It would be straightforward to compare the performance of such models trained on different L1s to humans, using the same methods described here. Outside of large language models, there are not many models that can be applied to wide swaths of language, but they do exist O'Donnell (2015); Abend, Kwiatkowski, Smith, Goldwater, and Steedman (2017) and could be elaborated to consider L1-L2 effects.

Another more data-driven approach would be to compare the distribution of language-specific morphosyntactic patterns in the L1 and L2. For instance, Römer and Yilmaz (2019) found that Turkish learners of English overuse the preposition *in* due to its role as a translation equivalent to several lexico-grammatical items in Turkish; these learners also overuse the *exist in N* construction and underuse the construction *there be*, potentially because the Turkish equivalent of the concept *exist*, 'var ol-', also means *there be*. This is a direction we are actively pursuing, though an interesting impediment is identifying texts that were written by native speakers. Many national corpora, for instance, include a mix of native and learner essays and do not explicitly distinguish the two, complicating analysis.

As with many papers, this one raises more questions than it answers. What it does, we hope, is paint a path towards answering those questions.

[1] For similar results using a very different method, see Johnson and Newport (1991).

References

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143.

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, 429–464.

Ahn, D., & Ferreira, V. S. (2024). Shared vs separate structural representations: Evidence from cumulative cross-language structural priming. *Quarterly Journal of Experimental Psychology*, *77*(1), 174–190.

Bermúdez-Margaretto, B., Gallo, F., Novitskiy, N., Myachykov, A., Petrova, A., & Shtyrov, Y. (2022). Ultra-rapid and automatic interplay between L1 and L2 semantics in late bilinguals: EEG evidence. *Cortex*, *151*, 147–161.

Berzak, Y., Nakamura, C., Flynn, S., & Katz, B. (2017, July). Predicting native language from gaze. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 541–551). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P17-1050/ doi: 10.18653/v1/P17-1050

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*.

Bley-Vroman, R. (1990). The logical problem of foreign language learning. *Linguistic analysis*, *20*(1-2), 3–49.

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.

Carlstrom, B., & Price, N. (2012-2014). *The Gachon Learner Corpus.*

Chen, T., & Hartshorne, J. K. (2021). More evidence from over 1.1 million subjects that the critical period for syntax closes in late adolescence. *Cognition*, *214*, 104706.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, *2*(3), 113–124.

Constantinescu, I., Pimentel, T., Cotterell, R., & Warstadt, A. (2025). Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, *13*, 96–120.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory (Wiley series in telecommunications and signal processing)*. USA: Wiley-Interscience.

Darģis, R., Auziņa, I., & Levāne-Petrova, K. (2018, May). The use of text alignment in semi-automatic error analysis: Use case in the development of the corpus of the Latvian language learners. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

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.

Di Nuovo, E., Bosco, C., Mazzei, A., & Sanguinetti, M. (2019). Towards an Italian learner treebank in universal dependencies. In *6th Italian conference on computational linguistics, clic-it 2019* (Vol. 2481, pp. 1–6).

DuPlessis, J., Solin, D., Travis, L., & White, L. (1987). UG or not UG, that is the question: A reply to Clahsen and Muysken. *Interlanguage studies bulletin (Utrecht)*, *3*(1), 56–75.

Dyer, W. E. (2017). *Minimizing integration cost: A general theory of constituent order*. University of California, Davis.

Epstein, S. D., Flynn, S., & Martohardjono, G. (1996). Second language acquisition: Theoretical and experimental issues in contemporary research. *Behavioral and Brain Sciences*, *19*(4), 677–714.

Futrell, R. (2024). Natural-language-like systematicity from a constraint on excess entropy. *Proceedings of the Society for Computation in Linguistics (SCiL)*, 336–337.

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.

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. (2003). The international corpus of learner english: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, *37*(3), 538–546.

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.

Hana, J., Rosen, A., Škodová, S., & Štindlová, B. (2010, July). Error-tagged learner corpus of Czech. In *Proceedings of the fourth linguistic annotation workshop* (pp. 11–19). Uppsala, Sweden: Association for Computational Linguistics.

Hartshorne, J. (2024). No evidence that age affects different bilingual learner groups differently: Rebuttal to van der Slik, Schepens, Bongaerts, and van Hout (2021). *Language Development Research*, *4*(1).

Hartshorne, J. K. (2022). When do children lose the language instinct? A critical review of the critical periods literature. *Annual Review of Linguistics*, *8*(1), 143–151.

Hartsuiker, R. J., & Bernolet, S. (2017). The development of shared syntax in second language learning. *Bilingualism (Cambridge, England)*, *20*(2), 219-234. doi: 10.1017/S1366728915000164

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in cognitive sciences*, *9*(5), 220–225.

Husain, S., & Vasishth, S. (2015). Non-projectivity and processing constraints: Insights from Hindi. In *Proceedings of the third international conference on dependency linguistics (depling 2015)* (pp. 141–150).

Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, ., Xu, X., & Glišić, I. (2021). *The Icelandic L2 Error Corpus (IceL2EC) version 1.1*. (CLARIN-IS)

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.

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.

Ivaska, I. (2014). The Corpus of Advanced Learner Finnish (LAS2): Database and toolkit to study academic learner Finnish. *Apples – Journal of Applied Language Studies*, *8*(3), 21–38.

Jarvis, S. (2017). Transfer: An overview with an expanded scope. *Crosslinguistic influence and distinctive patterns of language learning*, 12–28.

Johnson, J. S., & Newport, E. L. (1991). Critical period effects on universal properties of language: The status of subjacency in the acquisition of a second language. *Cognition*, *39*(3), 215–258.

Juffs, A., Han, N.-R., & Naismith, B. (2020). The University of Pittsburgh English Language Institute Corpus (PELIC) (1.0) [Data set]. *Zenodo*. doi: https://doi.org/10.5281/zenodo.3991977

Juola, P. (2008). Assessing linguistic complexity. *Language Complexity: Typology, Contact, Change. John Benjamins Press, Amsterdam, Netherlands*.

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).

Langacker, R. W. (1997). Constituency, dependency, and conceptual grouping. *Cognitive Linguistics*, *8*(1), 1–32. doi: doi:10.1515/cogl.1997.8.1.1

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.

Linzen, T., & Jaeger, T. F. (2014). Investigating the role of entropy in sentence processing. In *Proceedings of the fifth workshop on cognitive modeling and computational linguistics* (pp. 10–18).

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

Liu, Z., Eisape, T., Prud'hommeaux, E., & Hartshorne, J. K. (2022). Data-driven crosslinguistic syntactic transfer in second language learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Lozano, C. (2021). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, *0*(0), 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.

Malmasi, S., & Dras, M. (2015, May–June). Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 1403–1409). Denver, Colorado: Association for Computational Linguistics.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. Springer.

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).

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.

Montrul, S. (2011). Morphological errors in spanish second language learners and heritage speakers. *Studies in Second Language Acquisition*, *33*(2), 163–192. doi: 10.1017/S0272263110000720

Morgan-Short, K., & Ullman, M. T. (2022). Declarative and procedural memory in second language learning: Psycholinguistic considerations. In *The routledge handbook of second language acquisition and psycholinguistics* (pp. 322–334). Routledge.

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

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Osborne, T. (2014). Dependency grammar. In *The routledge handbook of syntax* (pp. 604–626). Routledge.

Pimentel, T., Meister, C., Wilcox, E. G., Levy, R. P., & Cotterell, R. (2023). On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, *11*, 1624–1642.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Preradović, N. M., Berać, M., & Boras, D. (2015). Learner corpus of Croatian as a second and foreign language. *Multidisciplinary Approaches to Multilingualism. Peter Lang*.

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., & 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.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of child language*, *14*(2), 201–209.

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

Römer, U., & Yilmaz, S. (2019). Effects of L2 usage and L1 transfer on Turkish learners' production of English verb-argument constructions. *Vigo international journal of applied linguistics*(16), 107-134. doi: 10.35869/vial.v0i16.95

Salamoura, A., & Williams, J. N. (2007). Processing verb argument structure across languages: Evidence for shared representations in the bilingual lexicon. *Applied psycholinguistics*, *28*(4), 627-660. doi: 10.1017/S0142716407070348

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, A. I., Kroll, J. F., & Diaz, M. (2007). Reading words in Spanish and English: Mapping orthography to phonology in two languages. *Language and Cognitive processes*, *22*(1), 106–129.

Schwartz, B. D., & Sprouse, R. A. (1996). L2 cognitive states and the full transfer/full access model. *Second language research*, *12*(1), 40–72.

Steedman, M. (2014). Categorial grammar. In *The routledge handbook of syntax* (pp. 670–701). Routledge.

Sun, K., & Wang, R. (2021). Using the relative entropy of linguistic complexity to assess L2 language proficiency development. *Entropy*, *23*(8), 1080.

Tenfjord, K., Meurer, P., & Hofland, K. (2006, May). The ASK Corpus - a Language Learner Corpus of Norwegian as a Second Language. In *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA).

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

Unsworth, S. (2010). First and second language development from a UG perspective. In *Language development over the lifespan* (pp. 25–45). Routledge.

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.

Wisniewski, K., Abel, A., Vodičková, K., Plassmann, S., Meurers, D., Woldt, C., ... Krivanek, J. (2018). *MERLIN written learner corpus for Czech, German, Italian 1.1.* (Eurac Research CLARIN Centre)

Wundt, W. M. (1910). *Völkerpsychologie; eine untersuchung der entwicklungsgesetze von sprache, mythus und sitte: Mythus und religion, 1. t* (Vol. 4). W. Engelmann.

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.

Yun, H., Li, W., Li, Z., & Hartshorne, J. K. (2023). Do children learn English more quickly when their native language is similar to English? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., … Znotiņš, A. (2024). *Universal dependencies 2.15.* (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University)