What can machine learning tell us about second language acquisition?

Acquiring a second language is a complex process that unfolds over years and proceeds differently for different aspects of language (Doughty & Long, 2008). Targeted studies of specific learner groups and particular linguistic phenomena have been informative, but it would be ideal to compare populations and phenomena directly in the same study — something that is not feasible with traditional methods because data-collection is too slow and costly. Here, we explore using machine learning and big data to fill this gap.

Our approach is to train a machine learning model to distinguish the morphosyntax of texts written by native speakers from those written by non-native speakers, then analyze what the models picked up on. This allows us to analyze the morphosyntactic behavior of individual learners as a whole, with the ability to discover patterns typical of non-native morphosyntax that may not be noticed by the human eye.

Data consist of 31,392 English essays by non-native speakers and 14,022 English essays by native speakers, sourced from 12 public corpora (Table 1). Following Berzak et al. (2014) and Liu et al. (2022), we characterized the morphosyntax used in each essay (patterns of syntactic relations, morphology, etc.) with 103 numerical features derived from an automatic universal dependency parser (version 2.9, (Zeman et al., 2021)). These include, for example, entropy (diversity) of parts of speech, number of distinct lemmas, average parse tree depth, and proportion of phrases with the verb preceding the object (see also Fig. 1). Effort was made to be comprehensive within the constraints of needing to derive the features automatically.

We considered several statistical classifiers such as logistic regression, decision trees, and support vector machines. Ultimately, random forest with class weight adjustment proved most accurate at distinguishing native from nonnative essays (accuracy = 0.94, f1 = 0.91, d-prime = 3.05). We confirmed that the model learned the differences in native and non-native writing, not just the differences between corpora: performance is still good within individual corpora that contain both native and nonnative essays (ICNALE: accuracy = 0.96, f1 = 0.88, d-prime = 3.43; BAWE: accuracy = 0.85, f1 = 0.91, d-prime = 1.70). Critically, model confidence was significantly correlated with writers' proficiency ($r=0.07$, $p<.001$), showing that the model's representations were sensitive to differences in proficiency among nonnative speakers.

Fig. 1 shows morphosyntactic features best distinguishing native and nonnative essays. We are currently conducting analyses that compare across first languages and across proficiency levels. We discuss limitations such as sample bias and methods to counteract.

Table 1 Descriptive statistics of native and learner corpora

Corpus	Genres	L1s	*N* of essays	*N* of tokens
TOEFL	Language test writings	Various L1s[2]	12,098	4,234,300
ICLE	Student essays	Various L1s[2]	9,480	6,654,960
CLC	Language test writings	Various L1s[2]	2,481	528,453
WriCLE-informal	Various types[1]	Spanish	1,059	835,551
ArabCC	Student essays	Arabic	957	223,938
WriCLE	Student essays	Spanish	706	710,942
BAWE	Students essays	English and Various L1s[2]	4,167	11,605,095
ICNALE	Argumentative essays	English and Various L1s[2]	4,626	1,207,386
OANC	Various types[1]	English	4,563	4,973,670
BNC	Various types[1]	English	4,167	2,133,504
ICE	Various types[1]	English	699	741,639
LOCNESS	Students exam essays	English	411	357,570

[1]Various types include blogs, emails, short autobiographical pieces, narratives, descriptions, and poems.
[2]Various L1s: Spanish, Japanese, Chinese, Korean, Arabic, French, German, Italian, Turkish, Hindi, Telugu, Thai, Greek, Portuguese, Polish, Tswana, Swedish, Russian, Hungarian, Indonesian, Persian, Lithuanian, Serbian, Norwegian, Macedonian, Bulgarian, Dutch, Finnish, Czech, Punjabi, Urdu, Catalan, and others (n<100).

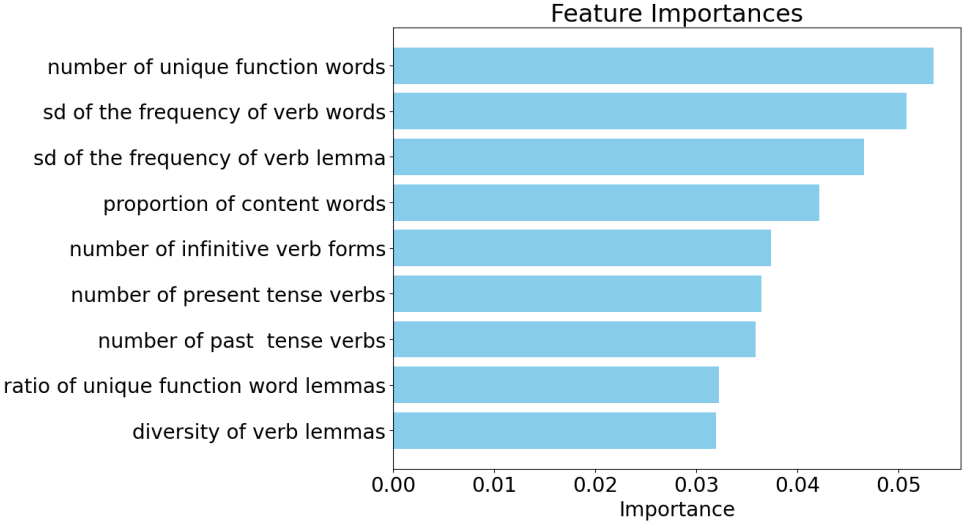

Figure 1 Morphosyntactic features most differentiating native and nonnative essays, based on random forest model feature importance

Reference:

Berzak, Y., Reichart, R., & Katz, B. (2014). Reconstructing Native Language Typology from Foreign Language Usage. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning (pp. 21-29).

Doughty, C. J., & Long, M. H. (Eds.). (2008). The handbook of second language acquisition. John Wiley & Sons.

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, No. 44).

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