What can machine learning tell us about second language acquisition?

Wei Li[1], Wenshuo Qin[1], Zoey Liu[2], & Joshua Hartshorne[3]

[1]Department of Psychology and Neuroscience, Boston College; [2]Department of Linguistics University of Florida; [3]MGH Institute of Health Professions

Introduction

Background

- Acquiring a second language is a complex process that unfolds over years and proceeds differently for different aspects of language.
- Some aspects of the language would be harder than others.
- 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.

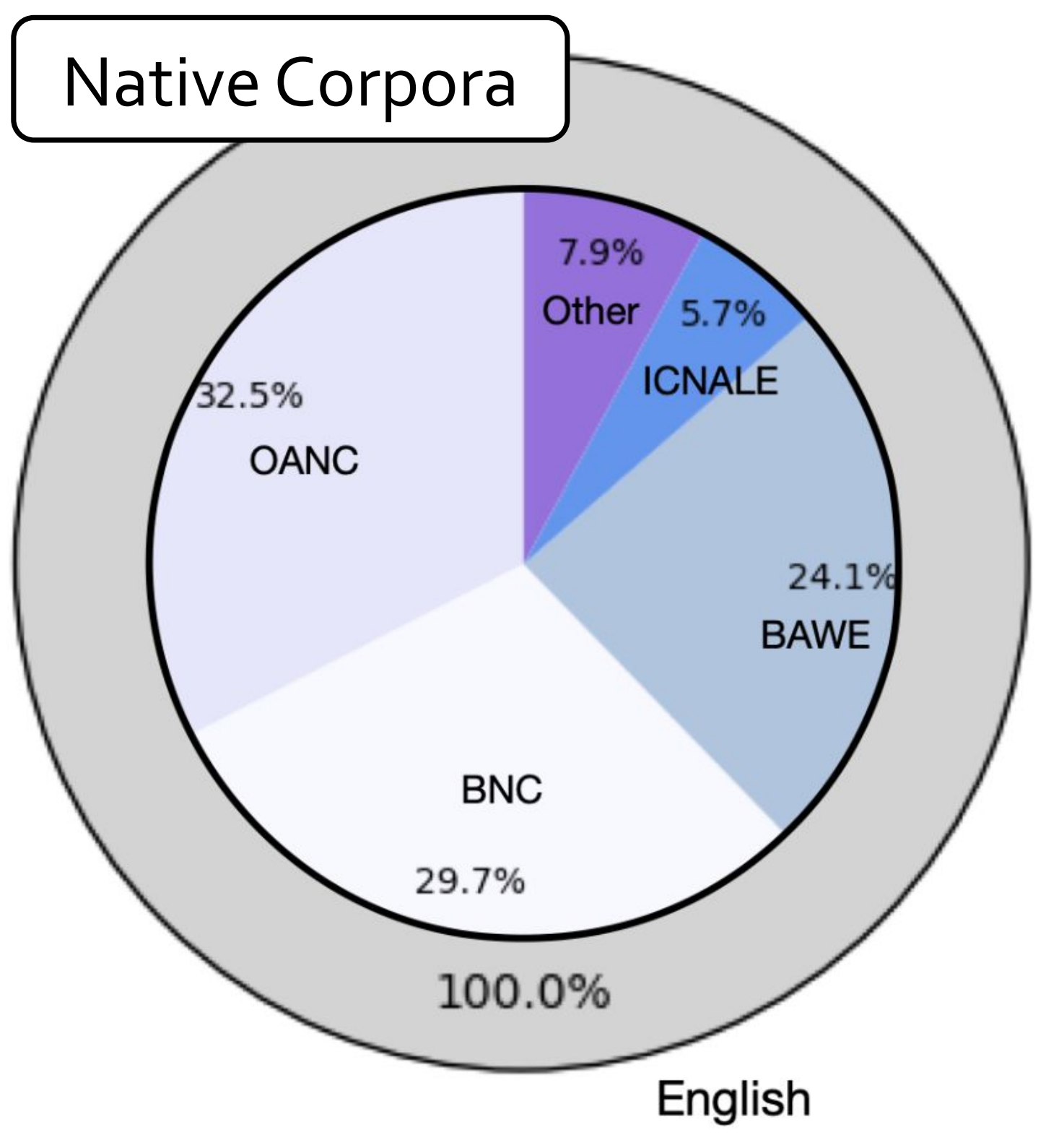
Current Study

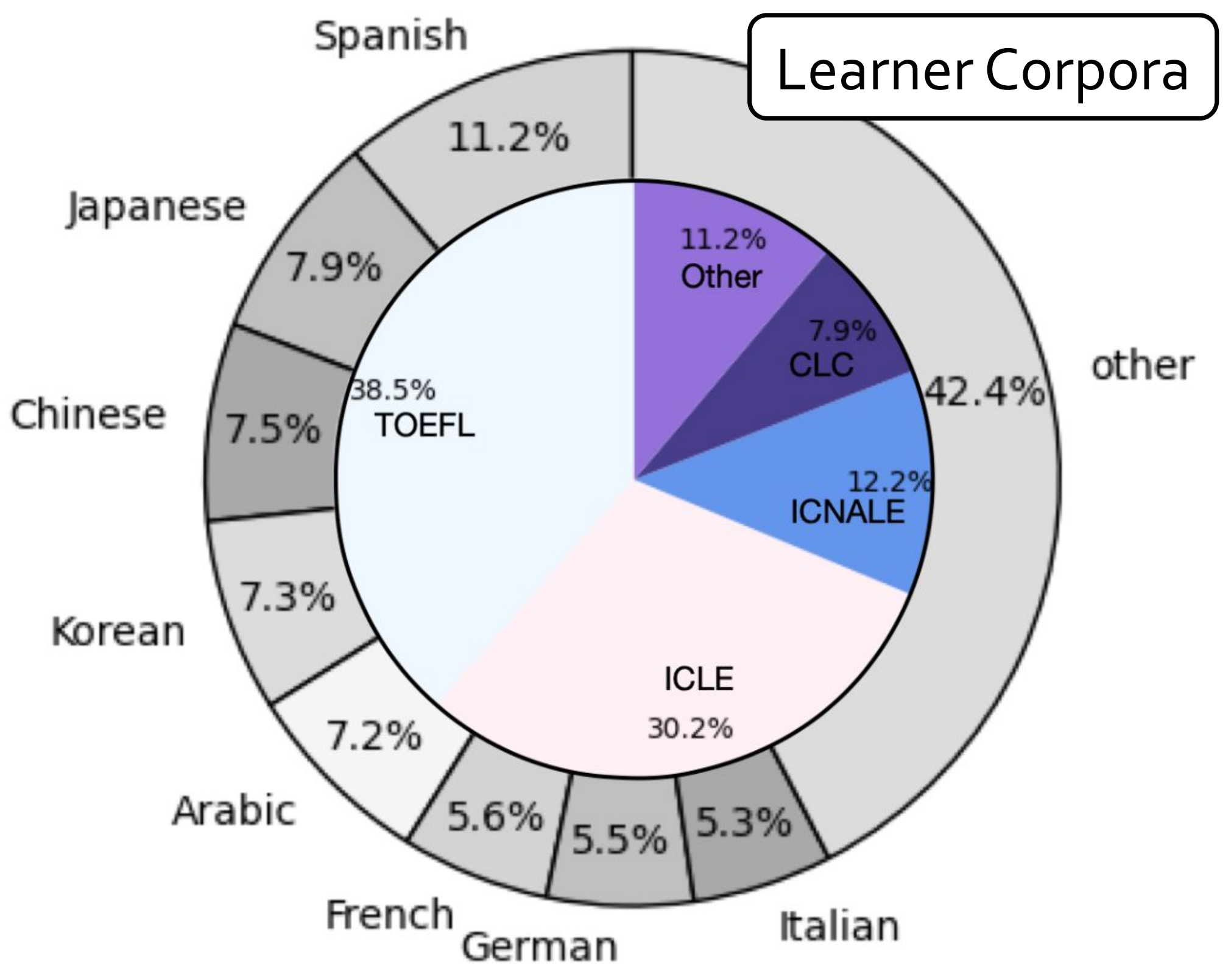
- Train a machine learning model to distinguish the morphosyntax of texts written by native speakers from those written by non-native speakers.
- Analyze the morphosyntactic behavior of individual learners as a whole, with the ability to discover patterns typical of non-native morphosyntax that may not be noticed by the human eye.

Method

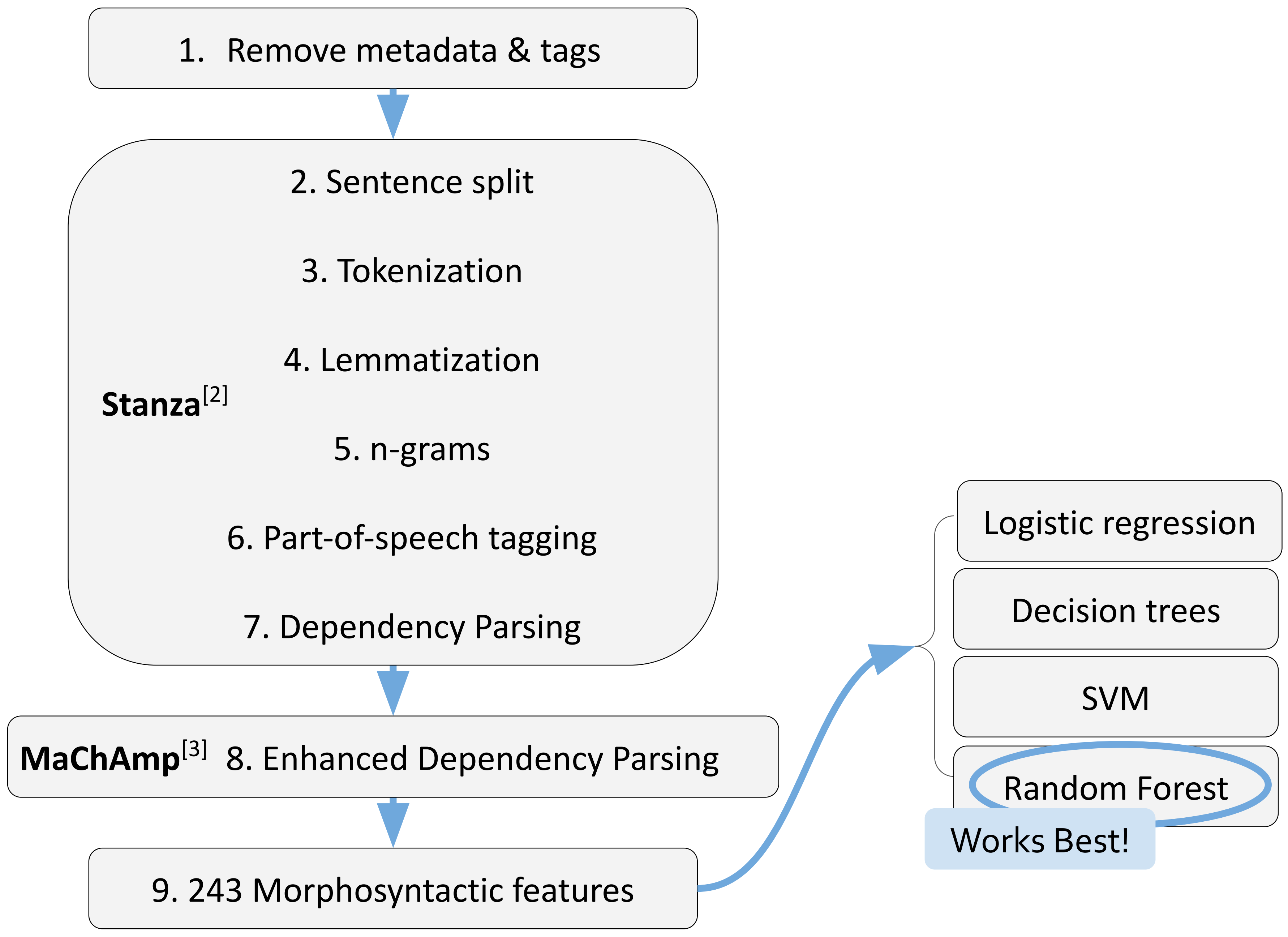
Training Dataset: corpora

	English Native Corpora	English Learner Corpora
Corpora Name	BAWE, LOCNESS, OANC, BNC, ICE, ICNALE	TOEFL, WriCLE, BAWE, ICNALE, CLC, ICLE, ArabCC
N of Essays	14,022	31,392
Genres	Newspaper, books, brochures, personal letters, university essays etc.	English class writings, exam writings, university essays, personal diaries, emails, blogs etc
L1s	English	52 first languages


Native Corpora


Learner Corpora

Data Processing



1. Remove metadata & tags

Stanza[2]

2. Sentence split

3. Tokenization

4. Lemmatization

5. n-grams

6. Part-of-speech tagging

7. Dependency Parsing

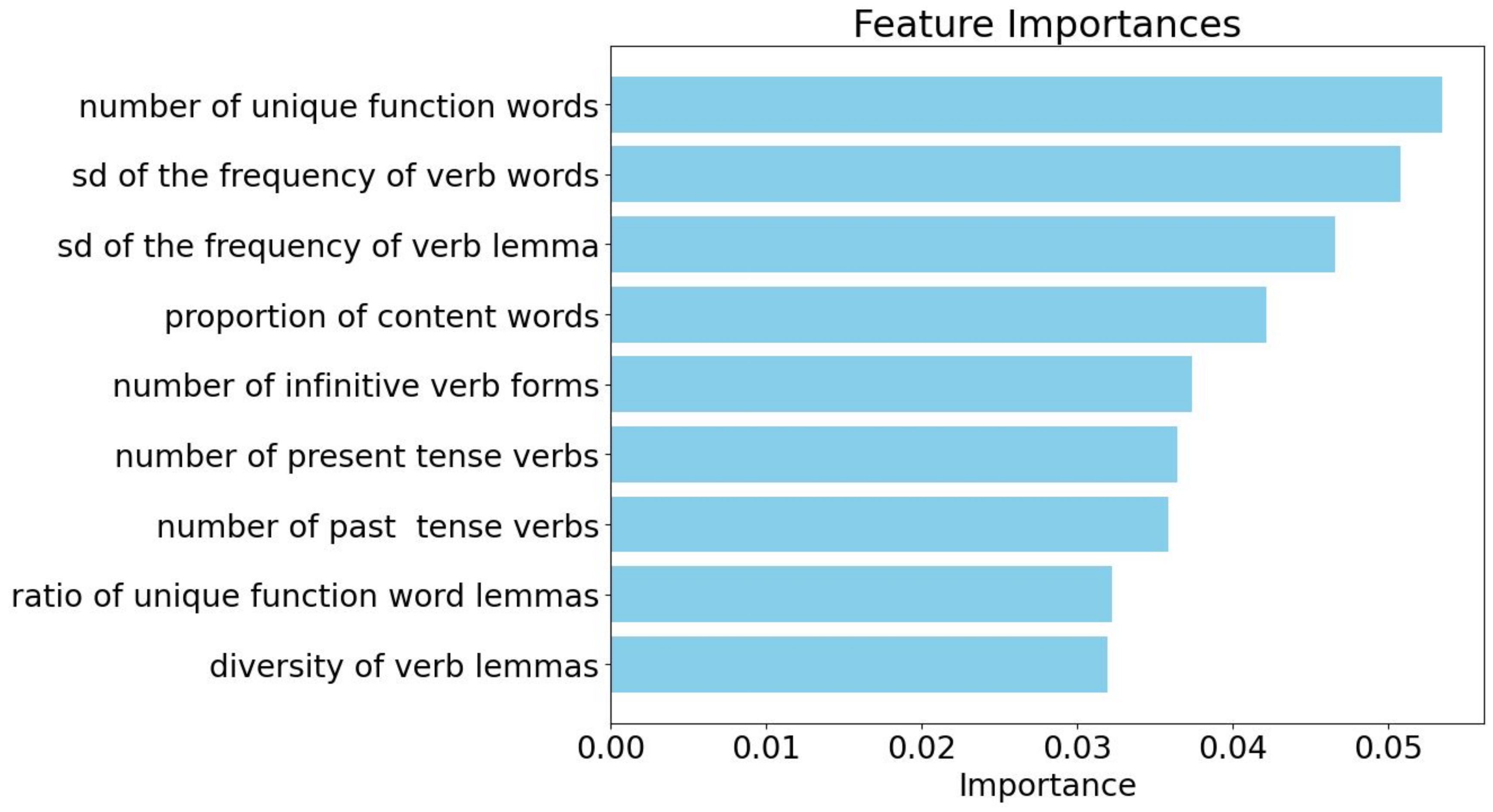
MaChAmp[3] 8. Enhanced Dependency Parsing

9. 243 Morphosyntactic features

Summary of Morphosyntactic Features

Category	Features
General	Average Sentence Length, Type-Token Ratio, Average Word Length, Lexical Density, Average Lemma Length
Word and Lemma	Function and Lexical Words/Lemmas: Types and Distributions
Verb	Analysis by Word, Lemma, Mood, Number, Person, Tense, Form, Valency, Aspect.
Auxiliary Verb	Analysis by Word, Lemma, Mood, Number, Person, Tense, Form: Entropy
Pronoun and Noun	Pronoun: Case, Number, Person, Type, Reflexives; Noun: Singularity; Detailed analysis of Demonstratives, Determiner Definiteness, Number Cardinality
Dependency Relations	Dependency Relation, Subordinate Dependency Relation
Syntactic Complexity	Average Dependency Length, Average Clause Length, Tree Depth Metrics
Miscellaneous	Verb Ratio, Adjective Degree, Prepositional and Conjunctive Lexical Diversity, Specific Lexical and Structural Configurations

Results & Discussion



Model Validation

- 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.

Interim Conclusion

- Function words and verb words are probably the more difficult for English learners to learn compared to other morphosyntactic features.
- We are currently conducting analyses that compare across first languages and across proficiency levels.

Reference

1. 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).
2. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.
3. Van Der Goot, R., Üstün, A., Ramponi, A., Sharaf, I., & Plank, B. (2020). Massive choice, ample tasks (machamp): A toolkit for multi-task learning in nlp. arXiv preprint arXiv:2005.14672.
4. Doughty, C. J., & Long, M. H. (Eds.). (2008). The handbook of second language acquisition. John Wiley & Sons.
5. 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).
6. 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.