

Automating Lexical Frequency Modeling in Chinese Using Interpretable Machine Learning

Lexical frequency is a fundamental property in Chinese language processing and learning, with norms from the Chinese Linguistic Data Consortium (CLDC) widely used in psycholinguistics and natural language processing. A newly released dataset of Chinese word production provides behavioral response times and rich linguistic annotations for thousands of words. This resource enables a novel perspective: instead of using frequency as an independent predictor, we ask whether frequency (CLDC_zipf) itself can be predicted from linguistic descriptors. This framing aligns with the theme of Chinese Language and Linguistics in the AI era, leveraging data-driven methods to connect theoretical constructs with corpus statistics.

We use the public dataset of Chinese word production with accompanying linguistic metrics. The target is CLDC_zipf (a Zipf-scale frequency from CLDC). Predictors include transparent, theory-driven features such as word length, phonological neighborhood size, character stroke counts, character and syllable frequencies, and age-of-acquisition. We train a suite of traditional ML models for regression: multiple linear regression, support vector regression (SVR), and tree-based ensembles (XGBoost, LightGBM, CatBoost). These models are chosen for accessibility and interpretability, explicitly avoiding black-box deep networks. We also apply SHAP (Shapley Additive Explanations) to the ensemble models to identify how each linguistic feature contributes to the frequency predictions.

Tree ensemble models achieved near-perfect predictions of lexical frequency. For example, XGBoost explained over 99% of the variance in CLDC_zipf ($R^2 \approx 0.994$) on held-out data, with extremely low error. Even the simpler linear and SVR models obtained very high R^2 (approximately 0.96–0.98), indicating that much of the frequency variation can be captured by additive linguistic factors. SHAP analyses confirm that predictors such as character-level frequency, word length, and age-of-acquisition, are the primary drivers of the model's predictions.

This work demonstrates that interpretable, automated ML methods can effectively model Chinese lexical frequency using linguistically grounded features. Unlike recent deep-learning trends, our transparent approach makes it clear which factors matter, bridging AI techniques and linguistic theory. We fill an important gap in the literature: prior studies on Chinese frequency mainly used hypothesis-driven regression analyses (often modeling how frequency affects reaction times), whereas we implement a fully automated predictive pipeline that uses rich annotations to predict frequency itself. Our findings show that traditional ML, combined with explainability tools like SHAP, provides powerful and accessible insights into Chinese lexicon structure. This interpretable AI approach paves the way for future Chinese linguistics and lexicography research, revealing how annotated lexical properties interact and advancing Chinese language science in the AI era.