Less is More: The Effectiveness of Compact Typological Language Representations

York Hay Ng[†], **Phuong Hanh Hoang**[†], **En-Shiun Annie Lee**[†,‡]
[†]University of Toronto, [‡]Ontario Tech University
{york.ng,fiona.hoang}@mail.utoronto.ca

Abstract

Linguistic feature datasets such as URIEL+ are valuable for modelling cross-lingual relationships, but their high dimensionality and sparsity, especially for low-resource languages, limit the effectiveness of distance metrics. We propose a pipeline to optimize the URIEL+ typological feature space by combining feature selection and imputation, producing compact yet interpretable typological representations. We evaluate these feature subsets on linguistic distance alignment and downstream tasks, demonstrating that reduced-size representations of language typology can yield more informative distance metrics and improve performance in multilingual NLP applications.

1 Introduction

The success of cross-lingual transfer in NLP often hinges on understanding relationships between languages (Lin et al., 2019). Resources such as URIEL (Littell et al., 2017) provide a large repository of linguistic features (typological, geographical, phylogenetic) for thousands of languages, encapsulating language properties in vector form. URIEL has been widely used to supply language features and vector distances to multilingual models (Lin et al., 2019; Adilazuarda et al., 2024; Anugraha et al., 2024). Khan et al. (2025) introduced **URIEL+**, extending URIEL's typological coverage to 4555 languages and addressing usability issues.

While URIEL+ represents a significant expansion in scope, its utility is constrained by the nature of its feature space. The typological set has grown to 800 features, many of which overlap or correlate strongly, as multiple sources contribute similar information (Littell et al., 2017; Khan et al., 2025). Furthermore, the data matrix remains sparse: after integrating five databases, 87% of values are still missing. Previous approaches have typically treated these features as equally informative, implicitly assuming uniform weight when computing language distances. This overlooks the high dimensionality and redundancy, which can introduce noise, risk model overfitting, and skew similarity metrics with uninformative signals.

Feature selection addresses these challenges by identifying and prioritizing the most statistically and linguistically informative features. This process mitigates the "curse of dimensionality" and can improve the separability of languages (Kohavi and John, 1997; Bellotti et al., 2014). In high-dimensional settings, correlation-based methods have been able to achieve a significant reduction in dimensionality, eliminating more than half of all characteristics, with minimal loss of predictive performance (Hall, 1999). Such focused selection can also act as a fast, scalable pre-processing step, substantially accelerating downstream tasks (Ferreira and Figueiredo, 2012). Applied in conjunction with missing-value imputation, this approach enhances data quality and generalization by focusing models on a core set of salient features (Liu et al., 2020).

We investigate optimizing the URIEL+ typological feature set with a pipeline of *feature selection* and *imputation* to improve its effectiveness for multilingual NLP, filling a key gap between the growing size of URIEL+ and the need for interpretable, task-driven feature sets. We move beyond the assumption of uniform feature importance and ask: *how can we produce compact, yet interpretable, typological language representations, and does this principled dimensionality reduction yield more meaningful language distances?*

Our contributions are: (1) the first principled framework for analyzing and selecting typological features, moving beyond treating all features as equally informative; (2) a robust imputation strategy with SoftImpute (Mazumder et al., 2010) to handle missing values within these compact representations; and (3) an empirical demonstration that these feature subsets improve linguistic distance alignment and downstream task performance.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25817–25828
November 4-9, 2025

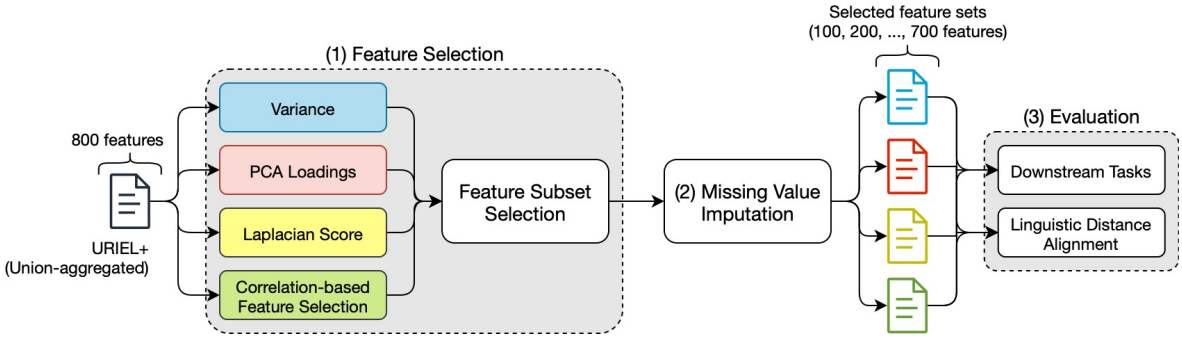

Figure 1: Overview of our optimization pipeline.

2 Methodology

Our optimization pipeline consists of three stages: (1) feature selection by four methods, including a novel approach of integrating language genetic data as class labels, (2) imputation of missing values, and (3) evaluation of the performance impact of each new feature subset. We perform our experiments on the URIEL+ *typological* dataset (Khan et al., 2025), a three-dimensional array of *languages* (4555), *features* (800), and *data sources* (14). We pre-process the dataset by taking union aggregation over sources (meaning a value is taken from any available source), producing a dataset of size $(4555, 800)$ containing *feature column vectors* $[\mathbf{f_0}, ..., \mathbf{f_{799}}]$ where $\mathbf{f_i} \in \{0, 1\}^{4555}$.

2.1 Unsupervised Feature Selection

We first employ three unsupervised filter-based methods to learn feature subsets. We evaluate each feature individually based on the statistical properties of the data, without the presence of class labels (Bellotti et al., 2014). As we vary the subset size k across $\{100, 200, ..., 700\}$, we rank features by importance scores s_i, $i \in [0, 800)$, and pick the top-k features based on the following metrics:

Variance: Ranking features by their variance has been found to be effective in sparse datasets (Liu et al., 2005; Ferreira and Figueiredo, 2012). Variance serves as a proxy for informativeness: a feature with minimal variance is either predominantly present or absent across languages, thereby offering utility in distinguishing between languages. The variance of a *binary* feature i is defined as:

$$s_i^{Var} = \text{Var}(f_i) = p_i(1 - p_i) \tag{1}$$

where f_i is a random variable representing feature i and p_i is the proportion of languages possessing feature i.

Principal Component Analysis (PCA) Loadings: Principal Component Analysis (PCA) identifies directions of maximal variance in the data via orthogonal principal components (PCs) (Jolliffe, 2014). Each PC is a weighted combination of the original features, where the weight assigned to a feature is called its *loading*. Intuitively, features with high absolute PCA loadings are those that explain the underlying structure of the typological dataset. We therefore select features that contribute most strongly to these components to retain interpretability while reducing dimensionality (Guo et al., 2002), scoring feature i by its maximum absolute loading among all PCs:

$$s_i^{PCA} = \max_{0 \le j < n} |b_{j,i}| \tag{2}$$

where n is the minimum number of PCs needed to explain 95% of the dataset variance, and $b_{j,i}$ is the loading for the i-th feature in the j-th PC.

Laplacian Score: Laplacian Score is a graph-based criterion that ranks features by how well they preserve locality in the data manifold (He et al., 2005). The intuition is that meaningful features should vary smoothly across similar languages.

We construct a 5-nearest neighbour graph over the languages, using a heat kernel to assign edge weights based on feature-space distance. Given a graph G with weight matrix S, degree matrix D, and unnormalized Laplacian $L = D - S$, we evaluate each feature $\mathbf{f}_i$ by first centering it with respect to D, producing $\tilde{\mathbf{f}}_i$, and computing its score:

$$s_i^{LS} = \frac{\tilde{\mathbf{f}}_i^\top L \tilde{\mathbf{f}}_i}{\tilde{\mathbf{f}}_i^\top D \tilde{\mathbf{f}}_i}$$

This score measures how much the feature varies locally relative to its overall variance. Lower-scoring features are more stable across similar languages and thus more likely to encode intrinsic typological patterns. We rank features in ascending order.

2.2 Supervised Feature Selection

While URIEL+'s language representations are inherently label-less, we further propose a strategy of applying language family membership as class labels. Using URIEL+ phylogenetic vectors, we encode the *top-level language family* as a categorical class variable c, allowing us to directly investigate feature relevancy with respect to the class.

Correlation-based Feature Selection: We utilize a method inspired by Correlation-based Feature Selection (Hall, 1999), producing a feature subset which balances between feature-class relevance and inter-feature redundancy. A "hill-climbing" algorithm (Kohavi and John, 1997) is used to build our feature set: starting with an empty set, we iteratively score all remaining features and add the feature yielding the highest score to the set.

We measure a feature's relevance by how well it distinguishes between top-level family membership. Specifically, we use Mutual Information (MI), a quantity used in linguistics and feature selection to measure how informative a feature is about the class (Bickel, 2010; Vergara and Estévez, 2014). It can be expressed as:

$$I(f_i; c) \;=\; H(f_i) + H(c) - H(f_i, c)\,, \quad (3)$$

where f_i and c are random variables representing the feature and the class, respectively, and $H(\cdot)$ denotes Shannon entropy (Cover and Thomas, 2006).

To mitigate redundancy between features, we scale the relevance score by the sum of correlations between the candidate feature and all features in the current subset. We measure the Phi correlation ϕ between features, which is ideal for binary features. The resulting merit score assigned to feature i is defined as:

$$s_i^{CFS} = \frac{I(f_i; c)}{\sum_{j=1, j\neq i}^{n} \phi(\mathbf{f_i}; \mathbf{f_j})} \quad (4)$$

where n is the current subset size. In each iteration of the search algorithm, this score is recomputed and determines the best feature to select.

2.3 Missing Value Imputation

To address missing values in the feature subsets, we apply SoftImpute (Mazumder et al., 2010) to complete the language-feature matrix, as it was demonstrated to be the strongest imputation method in URIEL+ (Khan et al., 2025). SoftImpute fills unknown typological values in the reduced feature space by approximating a low-rank matrix, distributing information from observed entries to infer missing ones. This low-rank assumption helps preserve global structure while reducing noise.

3 Linguistic Distance Alignment

To assess how well typological distances derived from feature subsets reflect linguistic similarity, we evaluate the alignment between our derived distances with a known linguistics similarity measure. We consider a set of similarity scores between 28 language pairs, comprising 24 low-resource languages in the Isthmo-Colombian area, identified in a case study by Hammarström and O'Connor (2013). These scores, derived from a modified Gower coefficient G_d, measure typological distance while accounting for feature dependencies.

We compute angular distances between language vectors from each feature subset, and report Spearman correlation coefficients ρ between our derived distances and the known similarity score G_d [1]

3.1 Alignment Results

Feature Set	Best Subset Size	ρ w.r.t G_d
URIEL+ (Baseline)	–	0.260
Variance	500	0.295
PCA Loadings	700	0.292
Laplacian Score	300	0.358
Correlation FS	200	0.264

Table 1: Exemplar alignment results for each method, in terms of Spearman correlations between language distances from feature subsets and the linguistic similarity measure G_d.

All feature selection methods, at the right subset size, outperforms the full URIEL+ feature space in distance alignment with linguistic reality. Although no feature subset produced distances which significantly correlated with G_d at the $p = 0.05$ significance level, considering the small sample size, the stronger alignments demonstrate the improved accuracy of distance-based comparisons between languages despite utilizing smaller feature sets, and suggests that our method enhances typological differentiability between languages.

Notably, the Laplacian Score strategy achieves the strongest G_d alignment, with $\rho = 0.358$ ($p = 0.061$) when using only 300 features, underscoring its effectiveness at capturing manifold

[1] Full results can be found in Appendix C.

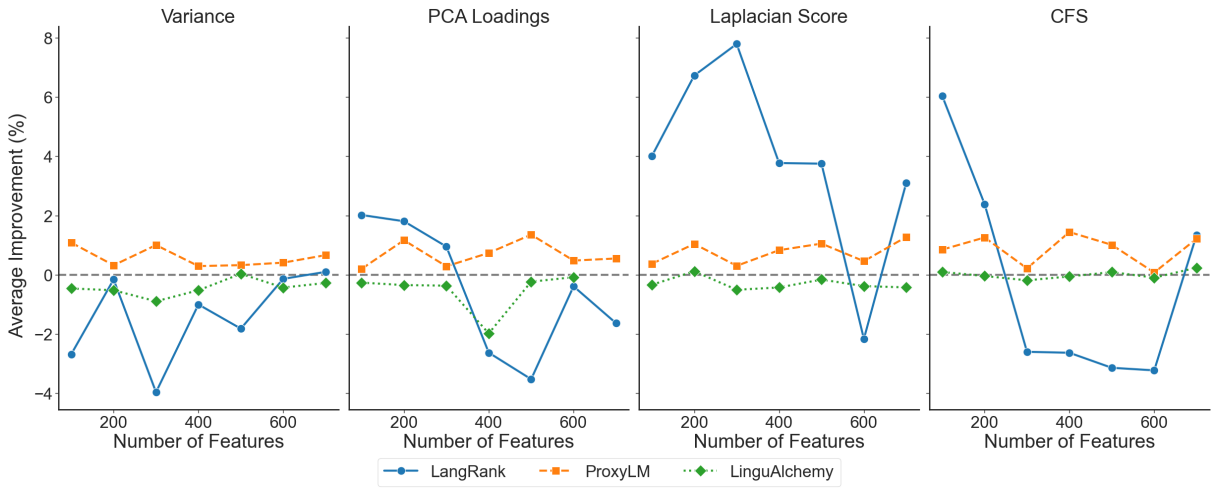

Figure 2: Average improvement (%) in downstream task performance: LANGRANK NDCG@3, PROXYLM RMSE and LINGUALCHEMY accuracy, compared to applying baseline URIEL+ distances.

structure. Simultaneously, in lower-dimensional spaces, the CFS strategy maintains similar alignment compared to URIEL+, highlighting its strength in selecting the most mutually informative features, while other methods excel in higher-dimensional spaces.

Despite the challenge of data sparsity, the consistent performance gains further indicate that our combined pipeline of feature selection and imputation can effectively trim noisy and redundant features, producing compact representations that generalize effectively for low-resource languages.

4 Downstream Task Performance

We test the utility of feature subsets on various NLP tasks which rely on typological features, focusing on application in multilingual tasks and models. [2]

4.1 Cross-Lingual Transfer Prediction

LANGRANK (Lin et al., 2019) is a framework for choosing transfer languages for NLP tasks requiring cross-lingual transfer based on language distances. For each feature subset we recompute angular distances between languages, grouped by type (syntax, morphology, etc.). These updated distances are supplied to LANGRANK to re-rank transfer languages across four subtasks: dependency parsing (DEP), machine translation (MT), entity linking (EL), and part-of-speech tagging (POS). We measure LANGRANK's ranking performance with average top-3 Normalized Discounted Cumulative Gain (NDCG@3).

[2]Detailed results and settings are shown in Appendix D.

4.2 Language Model Performance Prediction

PROXYLM (Anugraha et al., 2024) is a framework for predicting task performance of multilingual language models using typological distances. We replace the original distances with angular distances from each subset and train XGBoost regressors to predict performance of M2M100 (Fan et al., 2020) and NLLB (Team et al., 2022) on machine translation tasks MT560 (Gowda et al., 2021) and NusaTranslation (Purwarianti et al., 2023), evaluating with 5-fold root mean squared error (RMSE).

4.3 Language Model Linguistic Regularization

LinguAlchemy (Adilazuarda et al., 2024) introduces a linguistic regularization term that aligns text representations in language models with language vectors during training. While the original study used syntax vectors (i.e. comprising only syntactic features from the typological vector), we extract syntax vectors from each feature subset to replace the original vector. We subsequently measure the accuracy of mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019) on MasakhaNews topic classification (Adelani et al., 2023) and MASSIVE intent classification (FitzGerald et al., 2022) after aligning text representations.

4.4 Experimental Results

Figure 2 shows the average improvements in task performance across feature subset sizes (100–700) for each feature selection method, compared to URIEL+ distances and vectors.

For LANGRANK, the LS and CFS methods outperform the full feature set baseline at smaller subset sizes, with peak gains of +7.8% (size 300) and +6.0% (size 100) respectively. This highlights the utility of languages distances from these methods in choosing better transfer languages, while demonstrating how they excel in smaller subsets for capturing defining typological features. By contrast, Variance and PCA Loading strategies offer less improvement, indicating that feature variance and PC contribution alone are less effective at preserving structural information.

For PROXYLM, performance gains are smaller but consistent across all subset sizes and selection methods, with relative RMSE decrease of 0.2% to 1.4%. Crucially, no subset results in performance degradation. This suggests that, across subset sizes, our pipeline produces robust distances which contain sufficient information for predicting language model performance.

In comparison, when aligning text representations in language models to typological language vectors under the LINGUALCHEMY framework, performance varies only slightly, suggesting that the regularization loss is robust to feature reduction: as long as broad syntactic cues are preserved, smaller vectors provide signals comparable to the full set. This further supports the idea that compact representations suffice, with little benefit from carrying redundant features.

Overall, these results support our central hypothesis: principled feature selection and imputation reduces redundancy in typological datasets, improving the utility of distance values. Although different feature selection methods and subset sizes impact different downstream tasks uniquely, the fact that low-dimensional feature spaces from all four methods performs at least comparably to the 800-dimensional baseline in aiding LANGRANK and PROXYLM prediction suggests that reduced-dimensionality feature subsets better represent typological distance between languages, yielding improved predictive performance in downstream tasks while enabling more runtime-efficient distance computations.

5 Conclusion

We present a pipeline for producing compact, interpretable representations of typological features from the URIEL+ database. Through four feature selection strategies and imputation, we derived feature sets that improve the informativeness of typological distances even for low-resource languages, despite the dimensionality reduction of the feature space. Empirical results demonstrate their enhanced utility: new vectors improve typological comparability, alongside cross-lingual transfer and language model performance prediction accuracy, while reducing runtime costs. Our findings therefore suggest that *"less is more"* when modelling language relationships with typological features in multilingual NLP.

Limitations

Scalar distance. While our evaluation of feature subsets utilizes angular distances to quantify language similarity, any attempt to collapse complex linguistic relationships (encompassing syntax, phonology, and typology) into a single scalar distance inevitably oversimplifies. Different linguistic dimensions may require distinct representational strategies, and a unified distance metric cannot fully capture this richness.

Biases in feature selection. Each feature selection method inherently favors different properties: for instance, PCA favors features contributing to global variance, while CFS emphasizes local mutual relevance. As such, selected subsets may emphasize certain linguistic properties over others. A hybrid or ensemble-based selection strategy may yield more balanced representations.

Imputation techniques. We rely solely on SoftImpute for missing-value imputation. While it performs well under low-rank assumptions, alternative techniques (e.g., random forest models or autoencoders) could better capture linguistic relationships and offer complementary benefits, particularly for highly sparse languages or feature types.

Language coverage. In comparison to the 4,555 languages covered by URIEL+'s typological dataset, our evaluation focuses on a relatively small set of languages (19 in the alignment study, 105 in LANGRANK, 51 in PROXYLM and 63 in LINGUALCHEMY), with an emphasis on high-resource languages. While informative, this limits the generalizability of our findings across language families and resource levels. As seen in our results, our method achieves varying success across downstream applications. Future work should incorporate broader benchmarks, including diverse downstream tasks and typologically varied evaluation sets.

Ethics Statement

This work uses typological features which are derived from publicly available linguistic databases, and does not involve personal or sensitive data. Representing languages as feature vectors necessarily simplifies complex linguistic realities and may reflect biases in the source data, which users should keep in mind. While our contribution is primarily methodological and removed from direct applications, cross-lingual NLP carries both benefits (e.g., supporting low-resource languages) and risks of misuse. To support transparency and responsible use, we release our code publicly (Appendix E).

Acknowledgements

We would like to thank Mason Shipton and Jun Bin Cheng for their contributions, along with Giancarlo Giannetti, Aron-Seth Cohen and Bridget Green for the initiation of this study.

References

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Genta Indra Winata, Ayu Purwarianti, and Alham Fikri Aji. 2024. Lingualchemy: Fusing typological and geographical elements for unseen language generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3912–3928.

David Anugraha, Genta Indra Winata, Chenyue Li, Patrick Amadeus Irawan, and En-Shiun Annie Lee. 2024. Proxylm: Predicting language model performance on multilingual tasks via proxy models. *arXiv preprint arXiv:2406.09334*.

Tony Bellotti, Ilia Nouretdinov, Meng Yang, and Alexander Gammerman. 2014. Chapter 6 - feature selection. In Vineeth N. Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk, editors, *Conformal Prediction for Reliable Machine Learning*, pages 115–130. Morgan Kaufmann, Boston.

Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. 2017. Predicting native language from gaze. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 541–551, Vancouver, Canada. Association for Computational Linguistics.

Balthasar Bickel. 2010. Capturing particulars and universals in clause linkage: A multivariate analysis. In *Clause linking and clause hierarchy: Syntax and pragmatics*, pages 51–102. John Benjamins Publishing Company.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Thomas M Cover and Joy A Thomas. 2006. *Elements of Information Theory*, volume 1. John Wiley & Sons.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the*

North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Artur J Ferreira and Mário AT Figueiredo. 2012. Efficient feature selection filters for high-dimensional data. *Pattern recognition letters*, 33(13):1794–1804.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Q Guo, W Wu, D.L Massart, C Boucon, and S de Jong. 2002. Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, 61(1):123–132.

Mark A Hall. 1999. *Correlation-based feature selection for machine learning*. Ph.D. thesis, The University of Waikato.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.1. Max Planck Institute for Evolutionary Anthropology.

Harald Hammarström and Loretta O'Connor. 2013. Dependency-sensitive typological distance. In *Proceedings of the Workshop on Linguistic Distances 2013*, pages 329–352. Walter de Gruyter.

Xiaofei He, Deng Cai, and Partha Niyogi. 2005. Laplacian score for feature selection. *Advances in neural information processing systems*, 18.

Ian Jolliffe. 2014. *Principal Component Analysis*. John Wiley Sons, Ltd.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun A. Lee. 2025. Uriel+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2024)*, pages 6937–6952.

Ron Kohavi and George H John. 1997. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3125–3135.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 8–14.

Chia-Hui Liu, Chih-Fong Tsai, Kuen-Liang Sue, and Min-Wei Huang. 2020. The feature selection effect on missing value imputation of medical datasets. *applied sciences*, 10(7):2344.

Luying Liu, Jianchu Kang, Jing Yu, and Zhongliang Wang. 2005. A comparative study on unsupervised feature selection methods for text clustering. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 597–601.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322.

Yugo Murawaki. 2015. Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 324–334, Denver, Colorado. Association for Computational Linguistics.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning music helps you read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ayu Purwarianti, Dea Adhista, Agung Baptiso, Miftahul Mahfuzh, Yusrina Sabila, Samuel Cahyawijaya, and Alham Fikri Aji. 2023. Nusatranslation: Dialogue summarization and generation for underrepresented and extremely low-resource languages. *arXiv preprint arXiv:(coming soon)*.

Alex Rubinsteyn and Sergey Feldman. fancyimpute: An imputation library for python.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jorge R Vergara and Pablo A Estévez. 2014. A review of feature selection methods based on mutual information. *Neural computing and applications*, 24:175–186.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Related Work

Linguistic Feature Bases: URIEL and URIEL+. Littell et al. (2017) created the URIEL knowledge base to represent languages as vectors of linguistic features. These include typological features (drawn largely from the World Atlas of Language Structures and similar sources), phylogenetic features (language family indicators), and geographical coordinates, which are drawn from Glottolog (Hammarström et al., 2024). URIEL also defined distance metrics over these vectors, enabling calculation of inter-language distances (Littell et al., 2017).

URIEL+ (Khan et al., 2025) is an enhanced version that expands feature coverage and addresses limitations. It integrates additional databases (e.g. SAPhon for phonology, Grambank for morphology) to increase the number of features and languages covered. URIEL+ also introduces multiple imputation methods for missing data (including k-NN, MIDASpy, and SoftImpute) and allows customizable distance calculations. These changes improved downstream task performance and produced language distance estimates more aligned with linguistic reality (Khan et al., 2025).

Earlier work has sought to mitigate sparsity and redundancy in typological vectors. Murawaki (2015) used autoencoders to project WALS features into continuous latent spaces for phylogenetic inference, while Oncevay et al. (2020) fused typological vectors with task-learned embeddings to improve multilingual MT transfer.

However, prior to our work, the issue of redundancy and high dimensionality in typological vectors had not been explicitly addressed without sacrificing linguistic interpretability. In practice, researchers sometimes select subsets of features manually for specific tasks (Papadimitriou and Jurafsky, 2020; Zhang and Toral, 2019; Berzak et al., 2017), indicating that a one-size-fits-all feature set may be suboptimal.

Feature Selection and Reduction in NLP. Feature selection is a well-studied problem in machine learning, including NLP tasks where high-dimensional representations are common. Traditional approaches include filtering by statistical tests (e.g. chi-square for classification relevance) or greedy elimination of features with high pairwise correlation. In multilingual settings, Lin et al. (2019) learned to rank language features by importance for transfer learning. Other works have incorporated typological features into models via learned embeddings or regularization (Adilazuarda et al., 2024; Bjerva and Augenstein, 2018), implicitly performing a form of dimensionality reduction by focusing on the most informative typological aspects. Our approach explicitly reduces dimensions through trimming redundant features, allowing us to retain interpretability of features.

Missing Data Imputation. Missing feature values are prevalent in typological databases. In NLP, imputation has been explored to address the challenge of sparsity in linguistic databases such as WALS. Bjerva et al. (2019) proposed using probabilistic models and language embeddings to predict

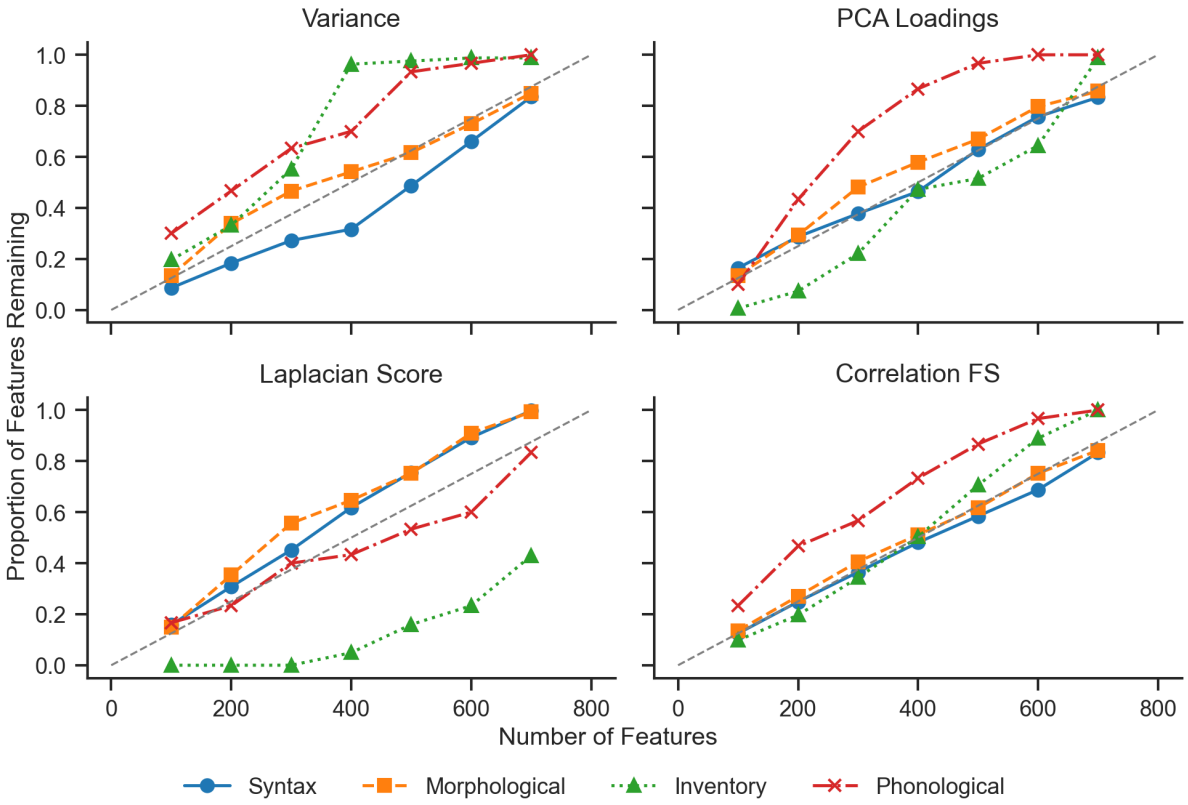

Figure 3: Proportion of features remaining in all four feature subsets, by feature type. Dotted line represents expected trend if selection is type-agnostic.

missing typological values, while Malaviya et al. (2017) used a neural machine translation model to predict missing values in the URIEL database. SoftImpute, proposed by Mazumder et al. (2010), is a matrix completion method that has shown the strongest performance in URIEL+ evaluations (Khan et al., 2025). We leverage SoftImpute to fill large portions of the URIEL+ matrix.

B Feature Selection Analysis

Through analyzing each feature subset, we can further investigate the linguistic or statistical properties favoured by different feature selection method.

Feature type. Analysis of selected feature types in each of the four subsets (Figure 3) further reveals structure in the optimized subsets. Inventory features, describing presence or absence of specific phonemes, were often excluded, especially by Laplacian and PCA-based methods, suggesting that they contribute less to preserving linguistic topology or transfer-relevant structure. On the other hand, phonological features were more frequently retained, particularly under PCA and CFS methods, implying phonological characteristics carry more discriminative power across languages in the context of multilingual NLP.

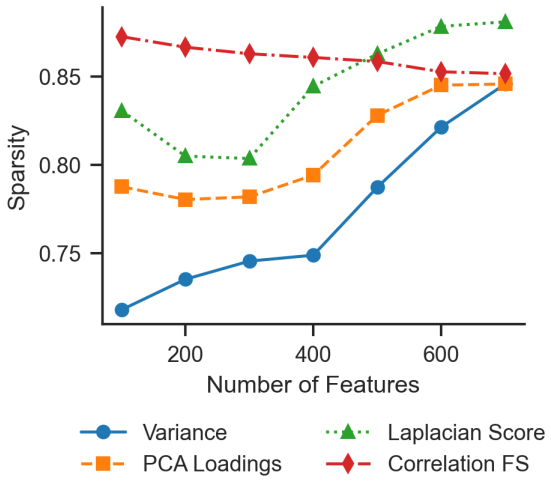

Figure 4: Average sparsity (proportion of missing values) of features across all four feature subsets.

Sparsity. While feature selection appears to be largely agnostic to sparsity, variance-based selection tends to favour less sparse features in smaller subsets, indicating that typological features with

Subset Size	Feature Selection Method			
	Var	PCA	LS	CFS
100	**0.289**	0.245	0.108	0.245
200	0.226	0.182	**0.339**	0.264
300	0.203	0.270	**0.358**	0.238
400	0.268	0.262	**0.319**	0.228
500	**0.295**	0.250	0.278	0.234
600	**0.292**	0.286	0.283	0.250
700	0.292	**0.292**	0.315	0.264
800	0.260			

Table 2: Correlation of distances derived from feature subsets with G_d for each selection method and subset size. Higher values indicate better performance. Best results for each subset size are highlighted.

broader coverage better balance presence and absence across languages. Correlation-based feature selection, conversely, favours *sparser* features, suggesting that these features are more relevant to language phylogeny.

C Linguistic Alignment

Table 2 shows detailed results of distance alignment with modified Gower coefficient G_d scores computed by (Hammarström and O'Connor, 2013), for all four feature selection methods: Variance (Var), PCA Loadings (PCA), Laplacian Score (LS) and Correlation-based Feature Selection (CFS). Each feature subset, at its best, yields stronger alignment compared to the baseline. The 24 low-resource languages studied were: Sambu, Cayapa, Paya, Bintucua, Cagaba, Ulua, Paez, Muisca, Huaunana, Boruca, Misquito, Quiche, Lenca, Cuna, Xinca, Camsa, Cofan, Colorado, Cabecar, Bribri, Catio, Bocota, Teribe, and Movere.

D Downstream Tasks

D.1 Experimental Settings

We ran LANGRANK with the same Glottocode replacements for some languages as detailed in Khan et al. (2025). We evaluated the performance of the trained LightGBM ranker by its mean Normalized Discounted Cumulative Gain score with $k = 3$, which measures the ranking quality of the three highest rankings, across cross-validation folds. Since LangRank uses linguistic features from a variety of datasets, we performed our experiment in two settings: (1) using only typological distance features and (2) using all features.

We used PROXYLM to train XGBoost regressors under the random test-split setting on model performance in both the English-centric and Many-to-many-language datasets, using the same regressor hyperparameters as in Anugraha et al. (2024). The performance of the regressor is determined by its root mean squared error (RMSE), which measures the difference between PROXYLM's predicted and actual values.

We replicated the LINGUALCHEMY pipeline used by Khan et al. (2025) for evaluating URIEL+'s syntax vectors. For each feature subset, we extracted syntax vectors by taking only the syntax features (features whose names begin with "S_") from the imputed, reduced-dimensionality vectors. For comparison with these syntax vectors, we applied URIEL+'s imputed syntax vectors as the baseline.

For evaluation, we computed average improvement by taking the average percentage change (or negative percentage change for PROXYLM experiments, as lower RMSE represents an improvement) in experiment scores, compared to the baseline, across all sub-tasks. No GPU was required to run either LANGRANK and PROXYLM, while all LINGUALCHEMY experiments were run on a single H100, utilizing 200 compute hours.

D.2 Detailed Results

Figures 5, 6, 7 and 8 show detailed experiment scores across sub-tasks for all feature subsets. For each feature selection method, evaluation scores across subset sizes are aggregated in a single distribution plot.

E Reproducibility and Licenses

All experiments were conducted with publicly available data. To support transparency and reproducibility, our code can be found at: `https://github.com/Swithord/URIELPlusOptimization`. In particular, our method uses the PCA implementation in `scikit-learn (v1.5)` (Pedregosa et al., 2011), along with the SoftImpute implementation in `fancyimpute (v0.7)` (Rubinsteyn and Feldman) with the mean initial fill method.

The URIEL+ database, along with evaluation frameworks PROXYLM, LINGUALCHEMY and LANGRANK, are licensed under CC BY 4.0.

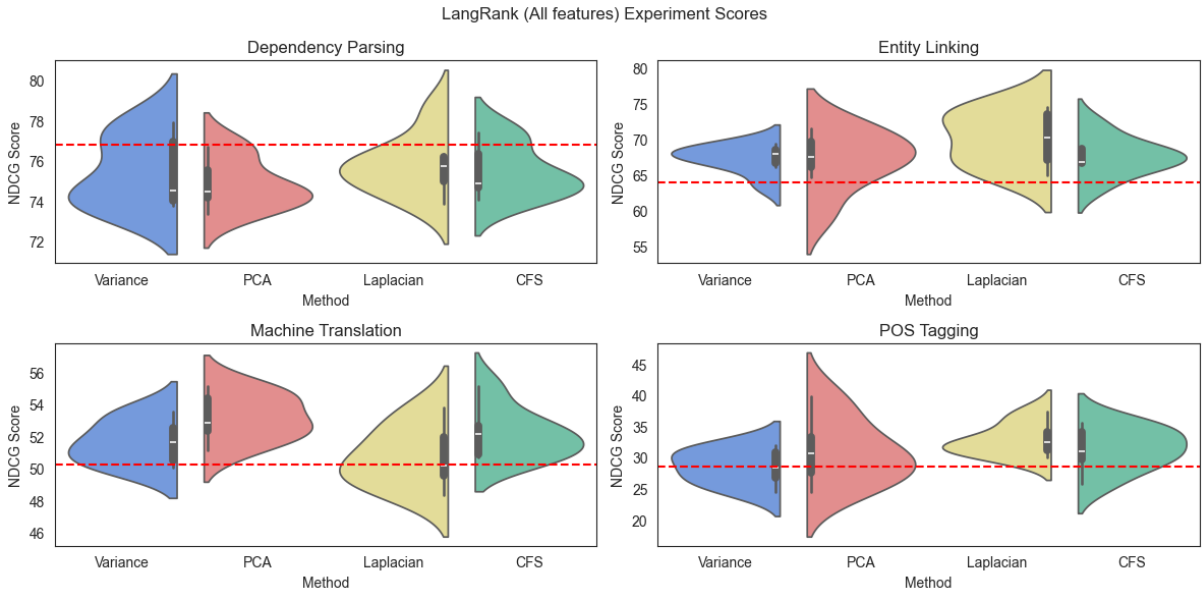

Figure 5: LangRank (using all features) NDCG@3 scores across each sub-task. Dotted line indicates baseline performance when training with URIEL+ distances. **Higher is better.**

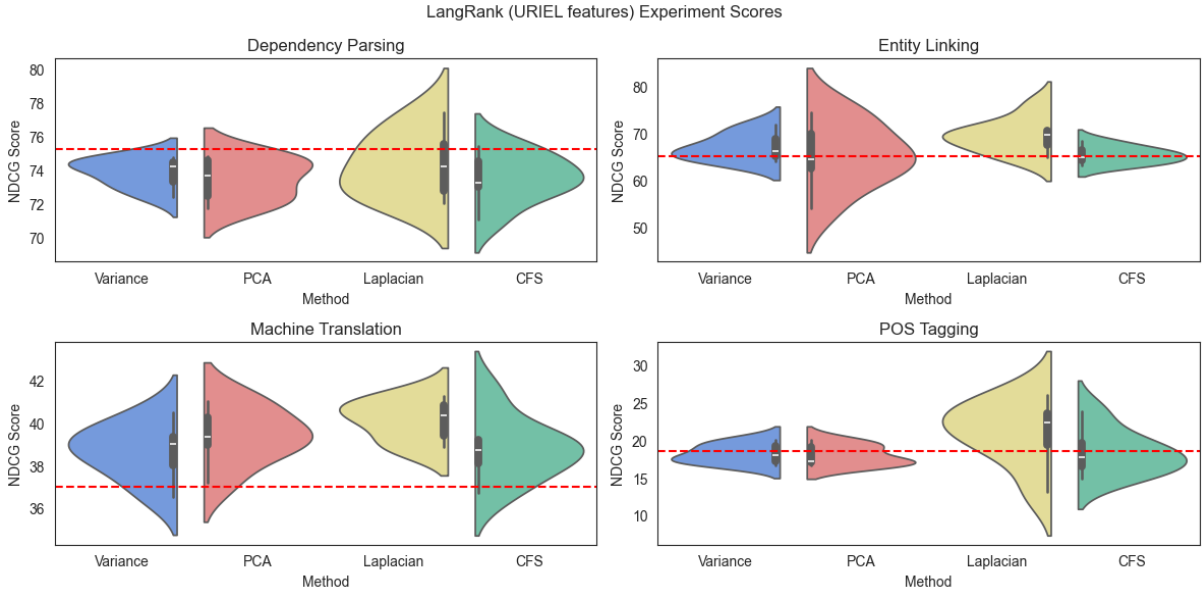

Figure 6: LangRank (using only URIEL features) NDCG@3 scores across each sub-task. Dotted line indicates baseline performance when training with URIEL+ distances. **Higher is better.**

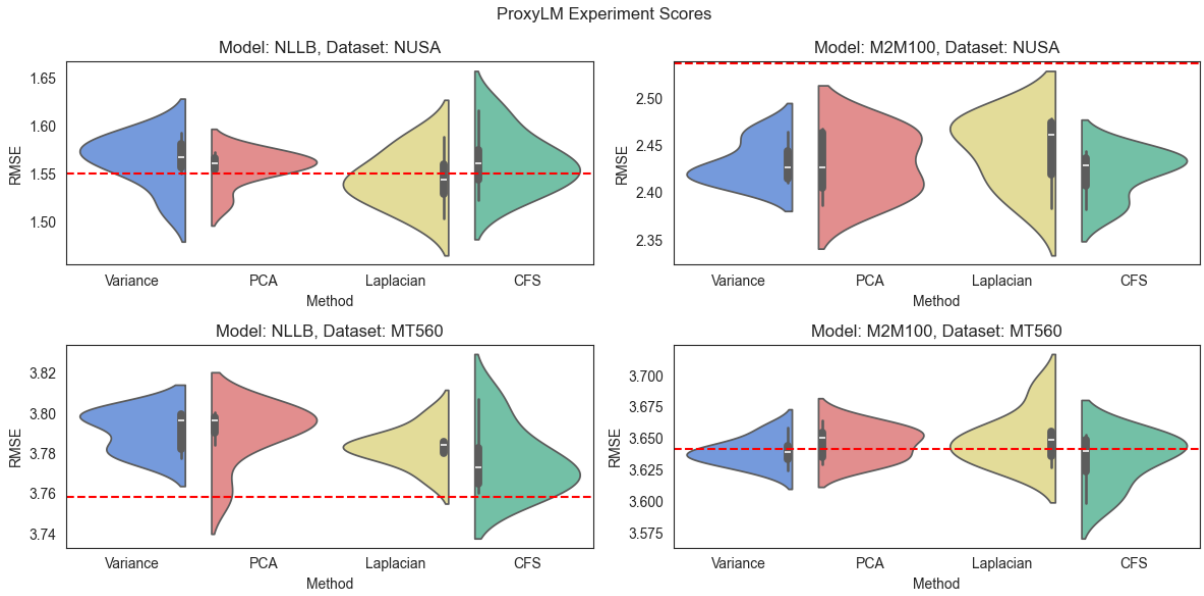

Figure 7: ProxyLM RMSE scores across each model and sub-task. Dotted line indicates baseline performance when training with URIEL+ distances. **Lower is better.**

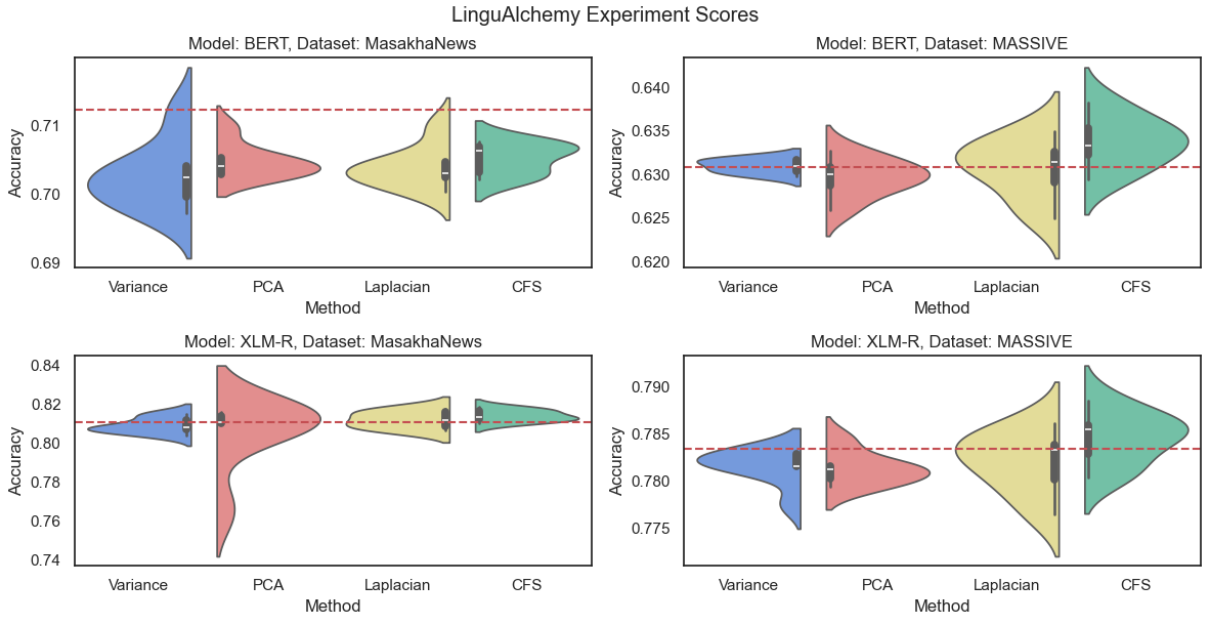

Figure 8: LinguAlchemy accuracy scores across each model and sub-task. Dotted line indicates baseline performance with URIEL+'s syntax vectors. **Higher is better.**