Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2026”

June 24–26, 2026

Towards Neural PIE Reconstruction from Heterogeneous Cognate Data

Konstantin Kalinowski and **Valentin Malykh**
ITMO University
49 Kronverksky Pr., bldg. A, St. Petersburg, 197101, Russia
costya5000@gmail.com valentin@maly.hk

Abstract

We construct a multi-source Proto-Indo-European (PIE) reconstruction dataset by aggregating IE-CoR, Kaikki/Wiktionary, Köbler, and EtymologyDB, with additional experiments on synthetic augmentation, validation, and phonetic input variants. We formulate reconstruction as sequence-to-sequence prediction from language-tagged cognate sets to PIE proto-forms and evaluate ByT5 together with data, representation, and architecture ablations. ByT5-Base with conservative Epitran-only phoneticization achieves EM = 0.216 on our PIE test set, a 59% relative improvement over the orthographic baseline. The results show that neural PIE reconstruction is feasible as a hypothesis-generation tool, but performance is driven mainly by data quality and representation fidelity rather than by larger models or aggressive augmentation.

Keywords: Proto-Indo-European, proto-form reconstruction, cognates, historical linguistics, neural reconstruction, ByT5

DOI: 10.29003/2075-7182-2026-24-163-170

Нейросетевая реконструкция праиндоевропейских форм по данным о когнатах из разнородных источников

Константин Калиновский и Валентин Малых
Университет ИТМО
197101, Россия, Санкт-Петербург, Кронверкский проспект, д. 49, лит. А
costya5000@gmail.com valentin@maly.hk

Аннотация

Мы формируем набор данных для реконструкции праиндоевропейских форм, объединяя IE-CoR, Kaikki/Wiktionary, словарь Кёблера и EtymologyDB, а также исследуем синтетическое расширение, валидацию и фонетические варианты входных данных. Задача формулируется как предсказание праформы по набору когнатов с языковыми метками. Мы оцениваем ByT5 и проводим абляционные эксперименты с данными, представлениями и архитектурами. ByT5-Base с консервативным преобразованием входных данных средствами Epitran достигает EM = 0,216, что на 59% превышает результат орфографической базовой модели. Результаты показывают, что нейросетевая реконструкция может использоваться для генерации гипотез, однако качество определяется прежде всего качеством данных и точностью представления, а не размером модели или агрессивным расширением данных.

Ключевые слова: праиндоевропейский язык, реконструкция праформ, когнаты, историческая лингвистика, нейросетевая реконструкция, ByT5

1 Introduction

Proto-Indo-European (PIE) is reconstructed rather than directly attested. In the classical comparative method, linguists compare cognate forms across daughter branches, identify regular sound correspondences, and infer the proto-form that best explains the observed reflexes. The task is difficult because the

evidence is uneven: some branches preserve conservative phonological traces, while others have undergone mergers, analogical leveling, or script changes. PIE notation also encodes distinctions that are only indirectly visible in descendants, such as laryngeals, palatovelars, labiovelars, and aspirated stops.

PIE reconstruction is also relevant beyond form prediction: lexical reconstructions are used in arguments about Indo-European chronology, homeland hypotheses, and cultural semantics, including wheeled-transport and hospitality vocabulary (Anthony and Ringe, 2015; Watkins, 2011).

This makes PIE reconstruction a useful stress test for neural sequence models. A model must process many languages, scripts, and annotation conventions, but the available training data is small by modern NLP standards. The recent IE-CoR release contains 4,981 curated cognate sets across 161 languages and 170 concepts (Anderson et al., 2025); this is substantial for historical linguistics, yet modest for neural reconstruction.

We ask whether a neural model can suggest useful PIE proto-forms from heterogeneous cognate evidence. The goal is not to replace specialist reconstruction, but to estimate how much recoverable signal is present in machine-readable cognate data and to provide a baseline for future work. Our contributions are:

- a multi-source PIE reconstruction benchmark built from IE-CoR, Kaikki/Wiktionary, Köbler, and EtymologyDB, with Starling parsed but excluded from the main branch because of notation and tagging inconsistencies;
- a ByT5-Base reconstruction baseline with controlled ablations over data expansion, language tags, phoneticization, tokenization, and transferred specialized architectures;
- an error and limitation analysis showing that progress depends more on data quality and representation fidelity than on larger models or aggressive synthetic expansion.

The strongest configuration, ByT5-Base with Epitran-only input conversion, reaches EM = 0.216 on the PIE test set, a 59% relative improvement over the orthographic baseline. The result is modest in absolute terms, but useful as a first reference point for neural PIE reconstruction on broad mixed-script data.

2 Related Work

Neural proto-form reconstruction has been studied in supervised, semi-supervised, and unsupervised settings. Ab Antiquo introduced an encoder-decoder baseline for proto-language reconstruction (Meloni et al., 2021). Later work reformulated the task with transformers and analyzed phylogenetic signal in learned representations (Kim et al., 2023). Reflex-prediction reranking was proposed in (Lu et al., 2024a), while CognateTransformer adapted an MSA-style transformer to proto-form reconstruction and cognate reflex prediction (Akavarapu and Bhattacharya, 2023). FeVeT introduced phonological feature vectors as segment representations (Wientzek, 2025). Semi-supervised reconstruction with DPD-BiReconstructor was proposed in (Lu et al., 2024b), and unsupervised neural reconstruction with monotonic alignment constraints was studied in (He et al., 2023).

Cognate detection and reflex prediction are closely related upstream tasks. The SIGTYP 2022 shared task established a benchmark for cognate reflex prediction across language families (List et al., 2022). Other work has addressed missing Romanian cognates and unattested Latin forms (Ciobanu et al., 2020), automatic cognate detection for Proto-Sabean reconstruction (Temesgen et al., 2025), and supervised cognate detection with CognateTransformer (Akavarapu and Bhattacharya, 2024).

Classical and statistical reconstruction remain important foundations. Probabilistic models of sound change were applied to ancient-language reconstruction in (Bouchard-Côté et al., 2013), and CRF-based Latin reconstruction was studied in (Ciobanu and Dinu, 2018). Related computational directions include neural sound-change prediction for phylogenetic inference (Chang et al., 2023) and neural language affiliation (Blum et al., 2025).

Our setting differs from most prior neural reconstruction work in scale and heterogeneity. Earlier systems were usually evaluated on typologically narrow datasets, often Romance, Germanic, Sinitic, or shared-task subsets with normalized IPA-like input. PIE reconstruction from broad Indo-European evidence requires mixed scripts, uneven attestation, and complex proto-form notation. Dataset resources

are therefore central: IE-CoR provides broad Indo-European cognacy coverage (Anderson et al., 2025; Heggarty et al., 2024), PILA offers aligned Proto-Italic and Latin data (Bothwell et al., 2024), and phonological studies of PIE features and laryngeals provide linguistic context for error analysis (Hartmann, 2019; Hartmann, 2021).

3 Data

3.1 Sources

We converted each entry into a unified sequence-to-sequence format: `input = [lang] form ...`, `output = PIE proto-form`. The main sources are IE-CoR (Anderson et al., 2025; Heggarty et al., 2024), Kaikki/Wiktionary PIE data (Kaikki.org, n.d.), Köbler (Köbler, 2014), and EtymologyDB. Starling (Nikolayev, 2012) was parsed but excluded from the main experiments because many entries contain branch tags, pseudo-tags, and legacy proto-form notation that are hard to normalize reliably.

Source	Rows	Mean cognates	Median	Max
IE-CoR	1,596	10.14	3	157
Kaikki	1,607	7.24	5	53
Köbler	399	1.67	1	19
Starling	3,177	4.15	3	15

Table 1: Parsed source resources before merging. Starling is shown for transparency but excluded from the main branch.

3.2 Dataset Variants

The base merged dataset (V1) combines IE-CoR, Kaikki, and Köbler: 3,602 rows split into 2,880 train, 361 validation, and 361 test examples. The V1 train split contains 22,830 tagged cognates, with mean 7.93 cognates per row and median 3.

The resulting inputs are heterogeneous not only by language but also by script. Typical segments include Latin-script forms such as `[eng] worm [lat] uermis [deu] Wurm`, Greek [grc] ἐρῠγγᾰνω, Devanagari Sanskrit [san] जघान, Armenian [xcl] դծաւմ, Persian in Arabic script [fas] ڕوغ, and transliterated fallbacks for Avestan and Hittite such as `[ave] aiha` and `[hit] ketu-`. This script diversity is one reason byte-level tokenization and conservative phoneticization are useful in our setting.

Synthetic augmentation was tested as a controlled data intervention, not assumed to be beneficial. For rows with fewer than 20 cognates, we generated up to 10 additional cognates per pass under strict formatting and language-uniqueness constraints. V2 uses one pass (+10), V3 uses two passes (+20), and V5 filters the synthetic additions with Wiktionary-guided validation. Validation used retrieved Wiktionary etymology snippets and batch LLM judgments to keep only additions marked `VALID`. This reduced the synthetic branch from 70,155 to 41,918 tagged cognates, but did not by itself improve reconstruction quality.

Separately, we expanded the non-synthetic base with EtymologyDB-derived material. This produced V6, an 8,836/361/361 split with 102,779 tagged cognates in train, mean 11.63 cognates per row, and 730 unique tags, of which 520 are ISO-3-compatible. From this split we created phonetic-input variants using Epitran (Mortensen et al., 2018), phonemizer/eSpeak-NG (Bernard and Titeux, 2021; eSpeak-NG Project, n.d.), and uroman (Hermjakob et al., 2018). The best phonetic variant, V10, uses conservative Epitran-only conversion.

The largest combined orthographic variant is V11:

$$\text{V11} = (\text{IE-CoR} \cup \text{Kaikki} \cup \text{Köbler}) + \text{validated synthetic}_{(+20)} + \text{EtymologyDB}.$$

After deduplication and split construction, V11 also uses an 8,836 train / 361 validation / 361 test split. We distinguish it from V10 because V11 tests whether validated synthetic evidence helps in orthographic input, while V10 tests representation quality through conservative phoneticization.

Variant	Main change	Train rows
V1	IE-CoR + Kaikki + Köbler	2,880
V3	V1 + two-pass synthetic augmentation	2,880
V5	V3 + Wiktionary-guided validation	2,880
V6	V1 + EtymologyDB	8,836
V10	V6 + conservative Epitran-only input	8,836
V11	V1 + validated synthetic + EtymologyDB	8,836

Table 2: Dataset variants used in the main experiments.

4 Methodology

We formulate PIE reconstruction as sequence-to-sequence prediction. The input is a concatenation of language-tagged cognates:

```
[lat] canis [grc] kyon [san] sva
```

and the target is the reconstructed PIE form:

```
*kuon-
```

We evaluate with four metrics: Exact Match (EM), Character Accuracy (CharAcc, defined as $1-\text{CER}$), Mean Edit Distance (MED), and Mean Normalized Edit Distance (MNED). DPD and FeVeT use native metrics that are not directly equivalent to these, so we report them separately when discussing transfer experiments.

Our main model is ByT5-Base (Xue et al., 2022). ByT5 operates on raw UTF-8 bytes, which is useful for mixed-script PIE data because it avoids out-of-vocabulary failures and preserves unusual proto-form symbols such as laryngeals, labiovelars, aspirates, and diacritics. We train with dropout 0.1, label smoothing 0.1, learning rate 5e-5, batch size 16, and early stopping with patience 5.

We compare ByT5 with subword-based T5-family models (mT5-Base, Flan-T5-Base, T5v1.1-Base), a larger byte-level model (ByT5-Large), and transferred specialized architectures: CognateTransformer (Akavarapu and Bhattacharya, 2023), DPD (Lu et al., 2024b), and FeVeT (Wientzek, 2025). These systems were designed for more normalized or typologically narrow settings, so their results on our data should be read as transfer baselines rather than native benchmark comparisons.

5 Experiments and Results

5.1 Main Results

Table 3 shows the key comparison. The V1 orthographic baseline reaches EM = 0.136. Conservative Epitran-only phoneticization on the expanded split (V10) gives the best overall result, EM = 0.216 and CharAcc = 0.490. The largest combined orthographic variant (V11) reaches EM = 0.183, making it the best non-phonetic input condition. CognateTransformer underperforms ByT5 after transfer to our data, especially when LingPy (List and Forkel, 2024) alignment is used on mixed orthographic strings.

Table 3: Main results on the PIE test set.

Configuration	EM	CharAcc	MED	MNED
ByT5-Base (V1)	0.136	0.449	3.02	0.365
ByT5-Base + conservative phoneticization (V10)	**0.216**	**0.490**	**2.84**	**0.344**
ByT5-Base + expanded orthographic data (V11)	0.183	0.452	3.10	0.371
CognateTransformer (LingPy on)	0.019	0.303	3.65	0.589
CognateTransformer (LingPy off)	0.055	0.317	3.52	0.564

Table 4 gives several predictions from the same ByT5-Base evaluation run. Correct cases tend to involve well-attested roots shared across several branches, while errors often involve laryngeal identity,

uncertain source annotations, or small suffix/length differences that still break exact match. Inputs are lightly transliterated where necessary for portable typesetting.

Table 4: Sample predictions from one ByT5-Base evaluation run. Inputs are truncated.

Status	Input	Target	Prediction
Error	[lit] sukti ...	*?*seuŭk-	*?*seuŭkk-
Error	[hit] ar- ...	*$h_3$er-	*$h_1$erH-
Correct	[hit] genu [lat] genu [grc] gony [san] janu ...	*gǧenu-	*gǧenu-
Error	[gae] troid ...	*?*treuŭd-	*treuŭd-
Error	[xto] wras [txb] warssam ...	*uŭer-	*uŭert-
Error	[bul] dobr [bel] dobry [ces] dobry ...	*d^heHb^h-	*d^heb^h-
Correct	[bul] vizdam [cat] veure [ces] videt [fra] voir ...	*uŭeiĭd-	*uŭeiĭd-
Error	[pol] toco ...	*temk-	*teiĭk-
Correct	[pal] manii [gaw] manug [pas] matu ...	*men-	*men-
Error	[gae] abhainn [hit] hapa- [xlu] hapa/i- ...	*h_2eb^h-	*$h_2$euŭ-i-

5.2 Ablations

Table 5 summarizes the main ablations. Removing language tags from V6 decreases EM from 0.183 to 0.161, showing that language identity is a useful signal. Subword-tokenized models are consistently weaker than byte-level ByT5, and ByT5-Large collapses in this low-resource setting. Synthetic augmentation is non-monotonic: +10 hurts, +20 recovers slightly, and validation alone reduces performance. Its best use is in V11, where validated synthetic additions are combined with EtymologyDB. For phoneticization, high-quality limited conversion is better than high-coverage noisy conversion.

Table 5: Selected ablations with ByT5-Base unless otherwise stated.

Configuration	EM	CharAcc	MED	MNED
V1 baseline	0.136	0.449	3.02	0.365
No language tags (V6)	0.161	0.448	3.09	0.378
No label smoothing	0.144	0.446	3.02	0.364
ByT5-Large	0.000	0.158	7.09	0.767
mT5-Base	0.070	0.220	4.50	0.550
Flan-T5-Base	0.031	0.307	3.73	0.470
T5v1.1-Base	0.042	0.299	3.85	0.481
+10 synthetic cognates (V2)	0.116	0.413	3.35	0.400
+20 synthetic cognates (V3)	0.144	0.440	3.24	0.381
+20 + validation (V5)	0.108	0.400	3.43	0.410
+20 + validation + EtymDB (V11)	0.183	0.452	3.10	0.371
Hybrid + uroman (V7)	0.175	0.474	3.04	0.367
Hybrid-full (V8)	0.177	0.462	3.03	0.365
Epitran-only (V10)	**0.216**	**0.490**	**2.84**	**0.344**

5.3 Language Transfer and Specialized Models

Leave-one-language-out experiments show that the model relies heavily on language-specific correspondences. Removing Latin reduces Latin-only EM from 0.144 to 0.061; removing Sanskrit reduces Sanskrit-only EM from 0.157 to 0.031; removing Greek reduces Greek-only EM from 0.141 to 0.018. Hittite is difficult even when present in training (EM = 0.054), likely because Anatolian evidence is sparse and cuneiform input is poorly handled by automatic phoneticization.

Transferred specialized models confirm that architecture alone does not solve the heterogeneous PIE setting. DPD performs better in semi-supervised than supervised mode on our V1 data (accuracy 0.069 vs. 0.017), but remains below ByT5. FeVeT on V7 reaches phoneme edit distance 3.56 and B^3 F1 0.307. These numbers are far below the original reports on narrower native benchmarks, which suggests a mismatch between their representation assumptions and our mixed-script, weakly normalized data.

6 Discussion and Limitations

The experiments support a data-centric interpretation. Neural PIE reconstruction is feasible, but current performance is constrained by evidence quality and representation fidelity. The best result comes from conservative Epitran-only phoneticization, not from maximal conversion coverage. Epitran changes only about one fifth of the segments, but those changes are high precision. Broader eSpeak and uroman fallbacks increase coverage while adding noisy romanization or rule-based G2P output, which weakens the reconstruction signal.

Synthetic augmentation is also limited. The V2 and V5 results show that generated cognates are not a reliable substitute for curated evidence. Validation removes many noisy additions, but filtering alone can reduce lexical diversity. The best orthographic result appears only when validated synthetic material is combined with EtymologyDB, so the conclusion is not that generation solves data sparsity, but that carefully filtered additions may help when anchored by external lexical evidence.

The error profile is linguistically plausible. Among targets containing laryngeals, the model recovers the full laryngeal set in only 31.1% of cases, with overall laryngeal presence F1 = 0.608. Deletion is the dominant failure mode, especially for h_1, H, and h_2. The model also confuses palatal and plain velars, over-generates aspiration, and often fails at the root level rather than making only feature-level errors. These patterns reflect genuine ambiguity in the daughter-language evidence: many branches merged relevant contrasts or preserve them only indirectly.

The practical use of the system is therefore limited but real. It should not be read as an automatic replacement for the comparative method. Performance also depends on evidence breadth: single-language examples reach only EM = 9.7%, compared with EM = 20.9% for multi-language cognate sets. A realistic use case is candidate generation or ranking: given a cognate set for a disputed item, the model can suggest plausible proto-form candidates for expert inspection. In this role, the system is useful precisely because its failures are measurable and can guide future dataset curation.

Future work should prioritize larger linguistically curated resources, stronger normalization of scripts and proto-form notation, and phonological representations designed for historical data rather than modern G2P coverage alone. Better integration of expert constraints may matter more than scaling model size.

7 Conclusion

We presented a neural baseline for PIE reconstruction from heterogeneous cognate data. ByT5-Base with conservative Epitran-only phoneticization achieves the best result, EM = 0.216, improving over the orthographic baseline by 59% relative. The main lesson is that representation quality matters more than coverage: precise partial phoneticization is more useful than aggressive conversion with noisy fallbacks.

The ablations also show that synthetic expansion is not automatically beneficial, subword tokenization is poorly matched to PIE notation, and transferred specialized architectures struggle outside their original homogeneous settings. The system is best understood as a hypothesis-generation tool for expert review, not as an autonomous reconstruction method. Progress will depend primarily on larger, cleaner, linguistically curated cross-script datasets.

This limitation is consistent with broader evidence that neural sequence models and transformer systems are sensitive to data scale and domain mismatch (Koehn and Knowles, 2017; Vaswani et al., 2017).

References

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab Antiquo: Neural Proto-language Reconstruction. In *Proceedings of NAACL-HLT 2021*. URL: https://aclanthology.org/2021.naacl-main.353/

Young-Min Kim, Kalvin Chang, Chenxuan Cui, and David R. Mortensen. 2023. Transformed Protoform Reconstruction. In *Proceedings of ACL 2023 (Short Papers)*. URL: https://aclanthology.org/2023.acl-short.3/

Liang Lu, Jingzhi Wang, and David R. Mortensen. 2024. Improved Neural Protoform Reconstruction via Reflex Prediction. In *Proceedings of LREC-COLING 2024*. URL: https://aclanthology.org/2024.lrec-main.762/

V.S.D.S. Mahesh Akavarapu and Arnab Bhattacharya. 2023. Cognate Transformer for Automated Phonological Reconstruction and Cognate Reflex Prediction. In *Proceedings of EMNLP 2023*. URL: https://aclanthology.org/2023.emnlp-main.423/

Tim Wientzek. 2025. Using Feature Vectors for Automated Phonological Reconstruction and Reflex Prediction. *Open Research Europe*, 5:174. DOI: https://doi.org/10.12688/openreseurope.20647.1

Andre He, Nicholas Tomlin, and Dan Klein. 2023. Neural Unsupervised Reconstruction of Protolanguage Word Forms. In *Proceedings of ACL 2023 (Long Papers)*. URL: https://aclanthology.org/2023.acl-long.91/

Elleni Sisay Temesgen, Hellina Hailu Nigatu, and Fitsum Assamnew Andargie. 2025. Cognate Detection for Historical Language Reconstruction of Proto-Sabean Languages: the Case of Ge'ez, Tigrinya, and Amharic. In *Proceedings of COLING 2025*. URL: https://aclanthology.org/2025.coling-main.496/

Alina Maria Ciobanu, Liviu P. Dinu, and Laurentiu Zoicas. 2020. Automatic Reconstruction of Missing Romanian Cognates and Unattested Latin Words. In *Proceedings of LREC 2020*. URL: https://aclanthology.org/2020.lrec-1.394/

Johann-Mattis List, Ekaterina Vylomova, Robert Forkel, Nathan Hill, and Ryan Cotterell. 2022. The SIGTYP 2022 Shared Task on the Prediction of Cognate Reflexes. In *Proceedings of SIGTYP 2022*. URL: https://aclanthology.org/2022.sigtyp-1.7/

V.S.D.S. Mahesh Akavarapu and Arnab Bhattacharya. 2024. Automated Cognate Detection as a Supervised Link Prediction Task with Cognate Transformer. In *Proceedings of EACL 2024*. URL: https://aclanthology.org/2024.eacl-long.58/

Kalvin Chang, Nathaniel Robinson, Anna Cai, Ting Chen, Annie Zhang, and David Mortensen. 2023. Automating Sound Change Prediction for Phylogenetic Inference: A Tukanoan Case Study. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*. URL: https://aclanthology.org/2023.lchange-1.14/

Frederic Blum, Steffen Herbold, and Johann-Mattis List. 2025. From Isolates to Families: Using Neural Networks for Automated Language Affiliation. *arXiv preprint arXiv:2502.11688*. URL: https://arxiv.org/abs/2502.11688

Alina Maria Ciobanu and Liviu P. Dinu. 2018. Ab Initio: Automatic Latin Proto-word Reconstruction. In *Proceedings of COLING 2018*. URL: https://aclanthology.org/C18-1136/

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229. DOI: https://doi.org/10.1073/pnas.1204678110

Cormac Anderson, Matthew Scarborough, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, et al. 2025. The Indo-European Cognate Relationships dataset. *Scientific Data*, 12(1):1541. DOI: https://doi.org/10.1038/s41597-025-05445-3

Stephen Bothwell, Brian DuSell, David Chiang, and Brian Krostenko. 2024. PILA: A Historical-Linguistic Dataset of Proto-Italic and Latin. In *Proceedings of LREC-COLING 2024*. URL: https://aclanthology.org/2024.lrec-main.1116/

Frederik Hartmann. 2019. Predicting Historical Phonetic Features using Deep Neural Networks: A Case Study of the Phonetic System of Proto-Indo-European. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 98–108. URL: https://aclanthology.org/W19-4713/

Frederik Hartmann. 2021. The phonetic value of the Proto-Indo-European laryngeals. *Indo-European Linguistics*, 9(1):26–84. DOI: https://doi.org/10.1163/22125892-bja10007

Paul Heggarty, Cormac Anderson, and Matthew Scarborough. 2024. Indo-European Cognate Relationships database (IE-CoR version 1.2). Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: https://doi.org/10.5281/zenodo.13304537 Project portal: https://iecor.clld.org

Kaikki.org. Proto-Indo-European machine-readable dictionary. Wiktionary-derived data; accessed June 7, 2026. URL: https://kaikki.org/dictionary/Proto-Indo-European/

Gerhard Köbler. 2014. *Indogermanisches Wörterbuch*, 5th edition. URL: http://www.koeblergerhard.de/idgwbhin.html

Sergei Nikolayev. 2012. Indo-European etymology. In *StarlingDB / The Tower of Babel*. URL: https://starlingdb.org/cgi-bin/response.cgi?root=config&basename=/data/ie/piet&table=0

Liang Lu, Peirong Xie, and David Mortensen. 2024. Semisupervised Neural Proto-Language Reconstruction. In *Proceedings of ACL 2024*, pages 14715–14759. URL: https://aclanthology.org/2024.acl-long.788/

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of LREC 2018*. URL: https://aclanthology.org/L18-1429/

Mathieu Bernard and Hadrien Titeux. 2021. phonemizer: Text to phones transcription for multiple languages in Python. *Journal of Open Source Software*, 6(68):3958. DOI: https://doi.org/10.21105/joss.03958

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box Universal Romanization Tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18. URL: https://aclanthology.org/P18-4003/

eSpeak-NG Project. eSpeak NG: open source speech synthesizer and rule-based grapheme-to-phoneme backend. URL: https://github.com/espeak-ng/espeak-ng

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39. URL: https://aclanthology.org/W17-3204/

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. URL: https://arxiv.org/abs/1706.03762

David W. Anthony and Don Ringe. 2015. The Indo-European Homeland from Linguistic and Archaeological Perspectives. *Annual Review of Linguistics*, 1:199–219. DOI: https://doi.org/10.1146/annurev-linguist-030514-124812

Calvert Watkins, editor. 2011. *The American Heritage Dictionary of Indo-European Roots* (3rd ed.). Houghton Mifflin Harcourt. ISBN: 978-0-547-54944-6. Catalog record: https://openlibrary.org/isbn/9780547549446

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, 10:291–306. URL: https://aclanthology.org/2022.tacl-1.17/

Johann-Mattis List and Robert Forkel. 2024. LingPy: A Python Library for Historical Linguistics, version 2.6.13. DOI: https://doi.org/10.5281/zenodo.597082 URL: https://lingpy.org