Constructing and Analyzing the Efficacy of a Proto-Germanic Cognate Dataset for Neural Proto-Form Reconstruction

Pablo Lanza Serrano
Department of Computer Science, University of Cambridge
pl486@cam.ac.uk

Abstract

We can learn what extinct languages sounded like through proto-form reconstruction by tracing back phonological and grammatical shifts. Meloni et al. (2021) and Kim et al. (2023) have developed supervised neural models that have achieved state-of-the-art results for reconstructing Latin and Middle Chinese, as the wealth of written records for these languages provide ample data to train their models on. We provide the first cognate dataset of Germanic languages for reconstructing Proto-Germanic, a language with no written records. Training a neural transformer model on variations of the dataset, reveals that not all descendant languages contribute equally and that removing sparse entries greatly improves performance, achieving comparable results with other datasets. Error analysis exposes a need for grammatical information to infer complex morphological patterns while analyzing the model's reconstructions shows that the model learns meaningful generalizations and is able to infer patterns of phonological change.

1 Introduction

Languages are constantly evolving as today's newest slang becomes tomorrow's standard speech. In a similar process to the evolution of species, a language's dialects can split-off and diverge until they become mutually unintelligible daughter languages. For the Germanic languages, this divergence happened around first and second centuries CE (van Coetsem and Kufner, 1972), with the ancestor language being termed Proto-Germanic (PGmc).

Proto-form reconstruction (PFR) is the task of deriving the words of an extinct language based on its attested descendants. Linguists are able to reconstruct these "proto-languages" by comparing cognates (words with a common origin) and deriving patterns of phonological change over these languages' evolution. Table 1 gives an example of how the PGmc phoneme /d/ has evolved in its descendant languages through systematic phonological shifts.

Eng.	Ger.	Swe.	Icel.	PGmc
mo**th**er	Mu**tt**er	mo**d**er	mó**ð**ir	mō**d**ēr
fa**th**er	Va**t**er	fa**d**er	fa**ð**ir	fa**d**ēr
wea**th**er	We**tt**er	vä**d**er	ve**ð**ur	we**d**ra
fo**th**er	Fu**tt**er	fo**d**er	fó**ð**ur	fō**d**rą

Table 1: Corresponding sounds in modern Germanic cognates and their Proto-Germanic (PGmc) ancestor.

To deduce these phonological patterns a large set of cognates is required. This is even more essential when employing machine learning methods, as current model architectures, such as recurrent neural networks (RNNs) and transformer models, require large amounts of data for accurate prediction. For supervised automatic PFR, we are therefore limited to well-resourced language families, or to those with a recent point of divergence where cognates and phonological shifts can be traced more easily.

The Germanic languages are an example of a family that has been extensively researched, with a large lexicon of reconstructed Proto-Germanic words. This makes PGmc an ideal candidate for reconstruction with machine learning methods. Nevertheless, modern machine learning architectures are notoriously data-hungry, placing a burden on the data collection to be as extensive and balanced as possible. This paper aims to answer whether a dataset of Germanic cognates can be obtained from publicly available data for accurate automatic Proto-Germanic reconstruction capable of inferring phonological patterns.

1.1 Contributions

We have compiled the first dataset of Germanic language cognates, each cognate-set matched with a corresponding PGmc ancestor word. The dataset was constructed from publicly available data on

Wiktionary, resulting in both an orthographic and phonological dataset.

We further used this dataset to train two state-of-the-art machine learning models for automatic PFR. The models trained on our full Germanic dataset under-perform when compared with other reconstructions of ancestral languages. However, the model successfully learns linguistically attested phonological patterns, while committing predictable errors with linguistic sources.

Finally, the PFR model is further trained on variations of the dataset with reduced sizes. These findings help us identify adverse entries in our dataset which we can remove to improve a naïvely constructed cognate dataset for automatic PFR. Ultimately, we achieve comparable results with other cognate datasets when training on a reduced dataset.

2 Related Work

Early attempts at proto-language reconstruction with machine learning methods aimed to explicitly capture the phonological changes that linguists use. Bouchard-Côté et al. (2013) use a probabilistic model that directly tracks sound changes, using a Monte Carlo inference algorithm. Their model is parameterized on a golden phylogeny of a language family (in their case the Oceanic languages) and takes an input of cognate-sets to produce an output of reconstructed ancestral forms, alongside a list of sound changes describing the evolution of the language family. Their implementation of automatic PFR is task inherently limited to ancestral forms produced with direct phonological changes, failing to capture changes in the phoneme inventories or morphological analysis.

A resurgence in the field was more recently brought about by Meloni et al. (2021) who take advantage of developments in machine learning and use a more sophisticated seq2seq RNN model for reconstructing Latin from modern languages. By extending and existing cognate dataset with data from Wiktionary, they are able to achieve state-of-the art results. This model was then used by Chang et al. (2022) to reconstruct Middle Chinese on another dataset constructed from Wiktionary.

A SIGTYP 2022 Shared Task (List et al., 2022c) challenges people to create cognate prediction models using the Lexibank dataset (List et al., 2022a). This dataset, while impressive in its breadth of languages, is insufficient for creating cognate-sets of sufficient size for automatic PFR. For example, Lexibank contains German has ~ 800 word entries, most of which missing a corresponding translation in other Germanic languages. Moreover, even when a translation is available the dataset does not ensure that they are truly cognates.

Finally, Kim et al. (2023) take the transformer model that has seen much success in other natural language processing applications to improve on Meloni et al.'s (2021) approach. Like Meloni et al. (2021), they adapt the standard encoder-decoder architecture to accommodate the structure of cognate datasets, where multiple daughter sequences correspond to a single proto-form sequence. They are able to improve upon Meloni et al.'s (2021) results, consistently having the best performance on most datasets they test.

3 The Dataset

The cognate dataset is comprised of entries consisting of a Proto-Germanic word and corresponding descendant words in daughter languages. It has been made publicly available under a Creative Commons 1.0 license[1]. Specifically, the dataset contains cognates from the following languages: Danish, Dutch, English, German, Gothic, Icelandic, and Swedish. These languages were selected for their large number of modern speakers corresponding to more entries in Wiktionary, and for their distinctive place in the Germanic phylogenetic tree (Figure 1). For example, Gothic is included as the only East Germanic language that we have a sizable lexicon for despite long being extinct.

The dataset was created out of the entries found in Wiktionary's category for Proto-Germanic lemmas[2]. Each entry had its Wiktionary page scraped, extracting the descendants corresponding to the languages in our dataset. Note that not every PGmc word has a descendant in each of its daughter languages. As such, each entry in the data set can have anywhere from 1 to 7 cognates (Table 2). Moreover, any capital letters and punctuation were also normalized across the cognates.

This Wiktionary data was used directly for the orthographic dataset. Like Meloni et al. (2021), we further generate a phonetic dataset of IPA phonemes by using the eSpeak library's[3] text-to-

[1] https://github.com/PLanza/Proto-Germanic-Cognates

[2] https://en.wiktionary.org/w/index.php?title=Category:Proto-Germanic_lemmas

[3] https://github.com/espeak-ng/espeak-ng

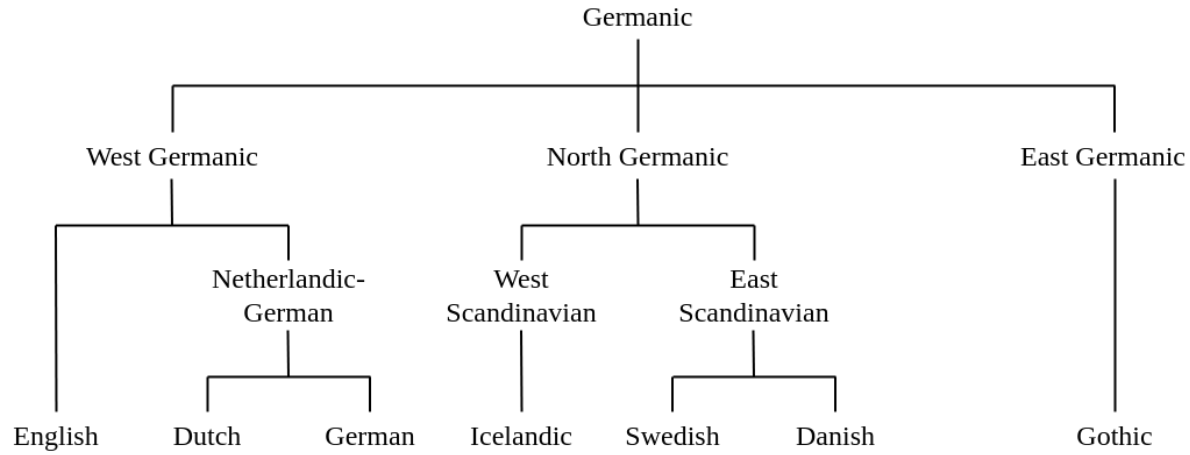

Figure 1: Phylogenetic tree of the Germanic languages included in our dataset.

Cognates	Sets
1	906
2	849
3	772
4	581
5	451
6	723
7	454
Total	4736

Language	Words
English	2663
Dutch	2508
German	2717
Swedish	2270
Danish	1955
Icelandic	3232
Gothic	1661
PGmc	4736
Total	21751

Table 2: Statistics of the data set. Cognates is the number of daughter cognates associated with a given Proto-Germanic word.

phoneme transcription tools. All syllable stress markers where removed from the output IPA transcription. This library was used for all the modern languages, but it unsurprisingly does not support Gothic nor Proto-Germanic transcriptions. Nevertheless, as pronunciations for these languages are reconstructed, they can be derived directly from the orthography, and so we can systematically transcribe the orthographic Gothic and Proto-Germanic into IPA phonemes according to Miller (2019) and Lehmann (2014) respectively.

The dataset was manually reviewed, checking for any errors in the automatic scraping and IPA generation processes. Certain PGmc words only made it into the modern languages as a morpheme rather than as a full word. For example, the Wiktionary page for the PGmc '*maisǭ' lists the English 'titmouse' as a descendant since the 'mouse' morpheme is indeed derived from '*maisǭ'. In such instances, we strip the duaghter word down to the descendant morpheme. Moreover, eSpeak's phoneme generation also struggled with short affixes, giving transcriptions of the letters' names (e.g. [ˌɛlwˈaɪ] for the adverbial suffix -ly). Finally, any entries that did not seem correct were checked against Kroonen (2013) to verify the etymological link.

The final dataset is comprised of 4736 cognate-set entries with a total 21751 words. This is over half the size of Meloni, et al.'s dataset of Romance cognates (8799) and substantially smaller than Chang, et al's Wikihan of Chinese cognates (21751). However, it is still substantially larger than any of List et al. (2022c) Lexibank language family cognate datasets. This demonstrates the difficulty of deriving large cognate datasets, even for well resourced languages like the Germanic languages. Furthermore, training (80%), validation (10%) and testing (10%) splits were generated for training the automatic PFR models.

4 Reconstructing Proto-Germanic

The purpose of this dataset is for reconstructing Proto-Germanic with machine learning methods. We will be testing the adequacy of our dataset on this task by using it to train two proto-language reconstruction models. These are Meloni et al. (2021) neural machine translation (NMT)-based RNN model, and Kim, et al.'s transformer-based model.

4.1 Preprocessing the Input

Both of the models used are seq2seq models that take a continuous sequence as input. However, our input data is a set of up to seven distinct daughter cognate words. To fit the data to the models, we concatenate the input words and encode them using

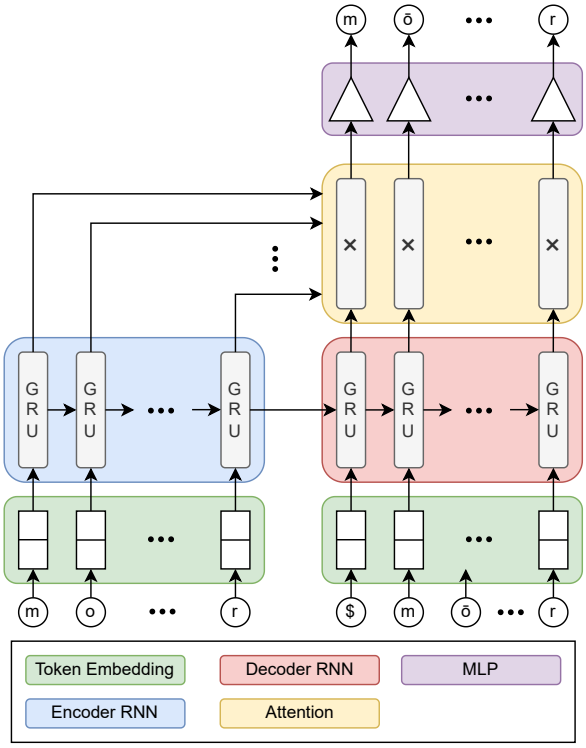

Figure 2: Component diagram of Meloni et al.'s (2021) RNN model. The bottom-left token sequence is the concatenated input descendant cognate-set. The top-right output is the Proto-Germanic prediction which is fed back into the decoder at the next time step. The '$' symbol represents a beginning of sequence token.

a character embedding table. This embedding is further augmented with an additive language embedding that imbues each character representation with information about which descendent language its corresponding word is from.

We also note that we do not always perform this encoding on individual characters, especially for the IPA inputs. The labialized velar consonants found in Gothic and Proto-Germanic (/k^{w}/, /g^{w}/, /x^{w}/) are composed of two Unicode characters but should be treated as a single phoneme token. This is similarly the case for phonemically lengthened vowels (e.g. in /aː/) and for the variety of other diacritics that the languages use in their orthography and IPA transcriptions. For these cases, we group the multiple characters together into single tokens.

4.2 Meloni et al.'s (2021) RNN model

Meloni et al. (2021) use a Gated Recurrent Unit (GRU) neural network model, also used for NMT tasks (Cho et al., 2014). They notice that like proto-word reconstruction, NMT generates words from some source language (daughter languages) to some target language (proto-language). Following a seq2seq architecture, they use GRU networks for encoding and decoding at the token level. The encoder takes a token from an input sequence containing the daughter language cognates, and outputs a contextualized representation that captures information about the relationships between tokens. As GRU networks are a kind of RNN, the encoder generates an output representation for each token in the input sequence, followed by and update the network's internal hidden state.

The decoder is also a GRU network that takes the encoder's final contextualized representation as its initial state. Where the encoder used the input daughter cognates sequence, the decoder now uses the previous time step's output as input for the current time step. Moreover, Meloni et al. (2021) use a dot-product attention mechanism to calculates the relevance that each part of the input has to the output (Bahdanau et al., 2014). So given our input consisting of the daughter cognates concatenated, we would imagine that the attention score for generating the first output token will match all the positions of the first token of each daughter word in the input sequence.

Finally, the decoder does not output the predicted proto-word's characters itself, rather it gives a vector representation of these characters that is passed on to a multi-layer perceptron (MLP). It is this perceptron that takes this representation and returns a probability distribution over the set of possible characters, taking an argmax to select the next character in the output sequence. This architecture is summarized in Figure 2.

4.3 Kim et al.'s (2023) Transformer-based model

Kim et al. (2023) similarly propose an econder-decoder model but using transformers instead, following recent state-of-the-art results in other natural language processing tasks. In addition to the character and language embedding performed on the input cognates, they further perform a positional embedding at the start. This preserves information about the positions of the tokens within each cognate, which is necessary as transformers' lack of recurrence means that they preserve no sense of sequential order.

The model's architecture follows closely that of Vaswani et al. (2017) (Figure 3). Transformers take the concept of attention and use it to construct the

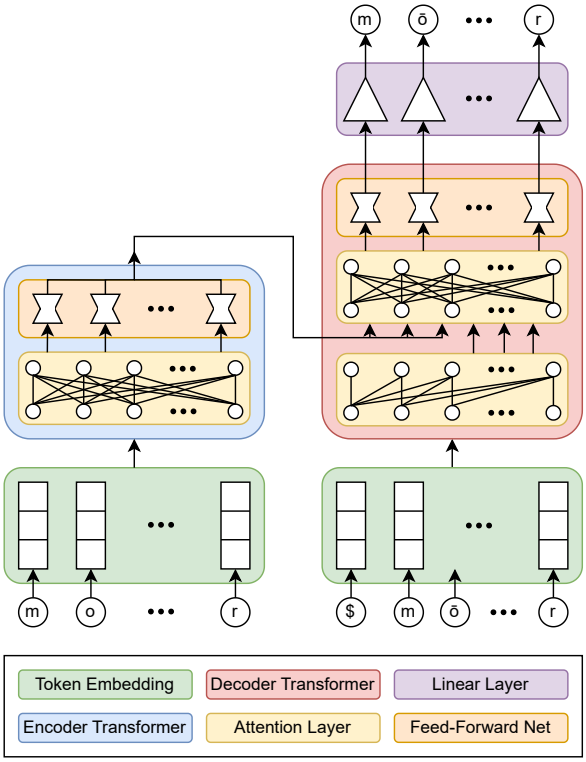

Figure 3: Component diagram of Kim et al.'s (2023) transformer model. The bottom-left token sequence is the concatenated input descendant cognate-set. The bottom-right input is the golden Proto-Germanic word, while the top-right output is the Proto-Germanic prediction. The '$' symbol represents a beginning of sequence token.

encoder and decoder, instead of solely as a means of communication between the two. Where an RNN looks at each token sequentially and remembers what parts of the input are relevant, attention layers can capture the relationships between the full input's tokens all at once. This gives transformers the advantage of parallelism. In the encoding phase, each input token sees the current representation of all the other tokens and updates its representation based on the information it has gathered about the others through an attention mechanism; these are the self-attention layers. Each self-attention layer — three in our model's case — is followed by a position-wise feed-forward network is applied, which introduces non-linearity to learn representations and patterns within a token's position, not just across positions through the attention layers.

The decoder similarly uses attention to generate each output token. However, the decoder cannot look ahead to future output tokens as they have not been generated yet. Instead, self-attention is calculated over the previous tokens. To take advantage of transformer's parallelism during training, we provide the model with the full gold output sequence but mask out future tokens when performing self-attention to not violate sequential generation. Then, decoder-encoder attention updates the output token's representation using the encoder's input representations in a similar manner to how attention was used in the RNN model. Like the encoder, a feed-forward network is applied after each encoder-decoder attention layer. Finally, similar to the RNN model, the decoder's output is passed through a linear layer that generates the output token's probability distribution.

Lang. Family	Kind	Descendants	Sets
Germanic	Ortho.	7	4736
	Phon.	7	4736
Romance	Ortho.	5	5419
	Phon.	5	5419
Sinitic	Phon.	39	804

Table 3: Number of descendants and cognate-sets for each dataset tested.

5 Experimentation

5.1 Baselines

We are aiming to determine whether our Germanic cognate dataset is suitable for proto-form reconstruction on the current state-of-the-art models. To do so, we compare against two baseline datasets that also derive their data from Wiktionary: Meloni et al.'s (2021) Romance orthographic and phonetic dataset, and Kim et al.'s (2023) Sinitic phonetic dataset used to reconstruct Middle Chinese. Note that this Romance dataset is not the full dataset that Meloni et al. (2021) use to evaluate their model, as part of it was constructed by Dinu and Ciobanu (2014) and is not publicly available.

5.2 Evaluation Metrics

Levenshtein distance (Levenshtein, 1965) is the standard metric for evaluating proto-form reconstruction systems which measures the character difference between two strings. Specifically, it counts the number of character insertions, deletions and replacements that are needed to convert the source string to the target string. Note that we consider multi-character tokens as single characters when calculating edit distance. However, this metric is insufficient to compare across datasets as, for example, many words in the Sinitic dataset are monosyl-

labic, and therefore on average shorter than those in the Germanic and Romance datasets. We therefore also use a normalized edit distance that divides edit distance by the length of the gold word. We also report on the accuracy of the reconstructions, i.e. the rate at which the edit distance is 0.

Finally, List (2019) notices that to properly evaluate the robustness of PFR systems, we should also evaluate their ability to infer patterns of phonetic shifts by measuring the structural similarity of predictions, rather than the surface level similarity. For example, PGmc nouns (singular number; nominative case) may have one of various endings, including '-az' and '-iz'. Say our system consistently predicts '-ok' and '-ek' for nouns with these stems respectively. Even though the endings would have a maximum edit distance of 2, the system has nevertheless successfully reconstructed a morphological pattern from the data that could have plausibly resulted from historical sound shifts. To address this, List (2019) proposes using B-cubed F-scores (Amigó et al., 2009) to measure the structural similarities of words.

6 Results

6.1 Initial findings

When comparing results across datasets the edit distance is not directly comparable, since as mentioned earlier, different datasets have different average word lengths. Instead we should focus on the NED as the comparable metric of how well a model trains on a given dataset. The results are also averaged over 10 training and testing sessions on both architectures. Nevertheless, we can see from Table 4 that our dataset performs the worst on all metrics, with a 19.5% difference in NED between our Germanic orthographic dataset and the worst baseline (Sinitic). We should also mention here, that due to time constraints we were unable to tune the models' hyper-parameters, and as such, used the same ones as Kim et al. (2023) for the Romance dataset. This is likely to have contributed to our dataset's under-performance.

On BCFS, the difference between the Sinitic and orthographic Germanic results is smaller at a 9.8% difference. This implies that the Germanic languages are descended from relatively more regular sound changes when compared with the Sinitic languages. It is also notable how Kim et al.'s (2023) transformer model performs better than Meloni et al.'s (2021) RNN model across the board.

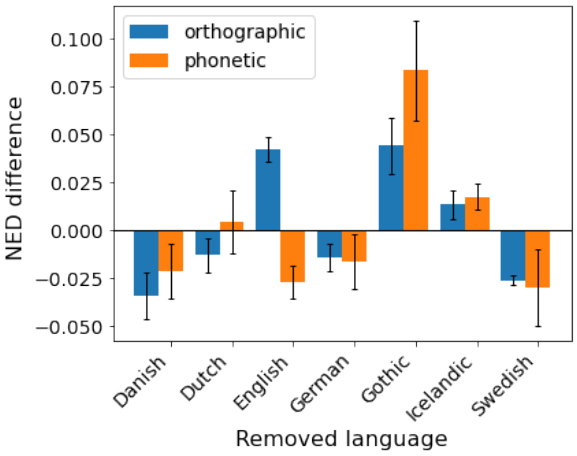

Figure 4: Changes in NED when training the transformer model on datasets with all the words of a given language removed. Data is averaged over 5 runs. Error bars indicate the standard deviation.

Similarly to the Romance dataset, both models perform better on the orthographic dataset. This is likely due to the fact that spelling tends to be more conservative than pronunciation and more strictly adheres to the historical phonological patterns. The patterns of phonological change would therefore be more obfuscated in the IPA data. Alternatively, the difference in performance could be simply due to the smaller alphabet of orthographic characters than of phonemes as the same letter could correspond to multiple allophones.

6.2 Contributions of languages

Given the scarcity of cognate datasets, it is worth investigating whether all languages contribute equally for automatic proto-form reconstruction. We do this by training the transformer model on our dataset, but removing all the cognates from each of the languages, one at a time. Intuitively, we expect the best results with languages spread across all branches of a given family tree, but in this experiment we will try to quantify whether this is actually true, and if there is any benefit in including, say German over Dutch despite their common origins. This information could be then used to guide decisions of which languages to include in future cognate datasets.

Surprisingly, Figure 4 shows that removing certain languages from the cognate-sets can improve the model's performance. The orthographic data seems to imply some connection between the language phylogeny and potentially superfluous data. German and Dutch, alongside Danish and Swedish

Dataset	Model	ED ↓	NED ↓	Acc. % ↑	BCFS ↑
Germ. (ortho.)	RNN (Meloni, et al.)	1.5962 ± 0.0418	0.2701 ± 0.0086	34.52% ± 0.22%	0.6590 ± 0.0040
	Transformer (Kim, et al.)	**1.5586 ± 0.0397**	**0.2635 ± 0.0076**	**37.55% ± 0.10%**	**0.6858 ± 0.0032**
Germ. (phon.)	RNN (Meloni, et al.)	1.7335 ± 0.2585	0.2762 ± 0.0143	36.82% ± 0.67%	0.6328 ± 0.0088
	Transformer (Kim, et al.)	**1.6551 ± 0.0539**	**0.2666 ± 0.0096**	**37.45% ± 0.30%**	**0.6474 ± 0.0048**
Rom. (ortho)	RNN (Meloni, et al.)	0.5958 ± 0.0083	0.0772 ± 0.0013	69.74% ± 0.23%	0.8913 ± 0.0016
	Transformer (Kim, et al.)	**0.5568 ± 0.0086**	**0.0724 ± 0.0013**	**71.15% ± 0.38%**	**0.8994 ± 0.0015**
Rom. (phon.)	RNN (Meloni, et al.)	0.9670 ± 0.0194	0.1229 ± 0.0020	52.09% ± 0.59%	0.8293 ± 0.0024
	Transformer (Kim, et al.)	**0.9027 ± 0.0194**	**0.1146 ± 0.0021**	**53.16% ± 0.66%**	**0.8421 ± 0.0029**
Sinitic	RNN (Meloni, et al.)	1.0720 ± 0.0536	0.2432 ± 0.0121	35.47% ± 1.40%	0.6747 ± 0.0166%
	Transformer (Kim, et al.)	**0.9814 ± 0.0437**	**0.2204 ± 0.0093**	**39.50% ± 3.02%**	**0.6971 ± 0.0102**

Table 4: Evaluation of training our Germanic cognates dataset and the baseline datasets on the two prot-form reconstruction models. Results are averaged across 10 runs, each trained using the same hyperparameters but on different random seeds.

are the two pairs of languages that diverged most recently from each other (Figure 1). These are also the ones that improve the model's performance when removed, implying that they could be superfluous as their closest relative already captures any relevant information passed down from PGmc. This trend, however, is weaker with the phonetic data as removing.

The outlying data point here is English, which goes from having a positive contribution in the orthographic experiments, to a negative contribution in the phonetic experiments. It seems that English's notoriously archaic orthography, despite frustrating many language learners, is useful for reconstructing PGmc. Contemporary English pronunciation has merged many sounds (e.g. the graphs 'ee', 'ea', 'ie', 'ei', 'e', 'i', 'y' could all represent the phoneme /i/ in English) making it harder for the historical sounds to be reconstructed.

The only two languages that don't result in an improvement in NED for either dataset when removed, are Icelandic and Gothic. This implies that they are necessary for accurate reconstructions of PGmc. Again, we can look to Figure 1 which shows these two languages splitting off from their relatives earliest. These languages could also be indispensable due to how Gothic was spoken closer to the time of PGmc than the modern languages, and how Icelandic is a very conservative being mutually intelligible with Old Norse in their written forms (Leonard, 2011). The archaic aspects of both these languages mean that they preserve many of the features present in PGmc which the model is able to pick up on and use in reconstruction.

6.3 Culling small cognate-sets

To attempt to improve the quality of our dataset, we propose culling the entries in the dataset with only a few daughter cognates. These entries may introduce spurious patterns that the model learns, leading to over-fitting. Moreover, using entries with many cognates allows the sound changes to be contextualized against the other daughter languages. This maps sound changes across the languages, allowing the model to more easily identify diverging sound shifts.

Figure 5 shows that on the orthographic data, the transformer model performs best when removing entries with only 1 descendant cognate, achieving a minimum NED of 0.2188. On the phonetic data,

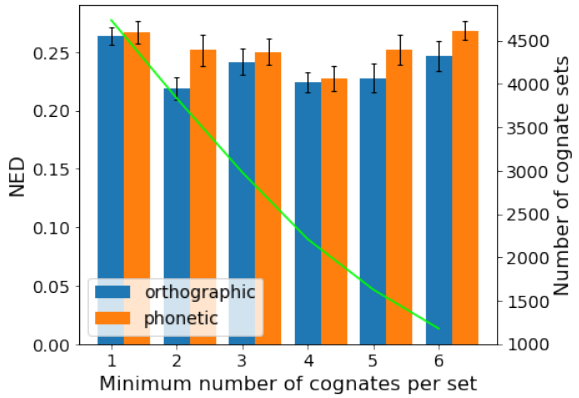

Figure 5: NED of the transformer model when trained on the Germanic datasets with entries containing a minimum number descendant cognates. The size of these reduced datasets is also plotted in green. NED is averaged over 5 runs. Error bars indicate the standard deviation.

the same model performs best when removing the entries with fewer than 4 cognates, with a minimum NED of 0.2275. This shows that even though we have reduced the size of the datasets, the model is better able to learn the patterns between cognates, resulting in better reconstruction. When compared to the baselines, our minimum NED is now below that for the Sinitic dataset. Furthermore, BCFS for the best performing reduced datasets is 0.7142 and 0.7000 for the orthographic and phonetic datasets respectively; above that of the Sinitic dataset. This shows that the reduced datasets allow the model to better infer structural patterns from the data as well.

Unfortunately, training on these reduced datasets does not generalize to testing on data with the sparse entries included. Attempts at training the transformer model with the sparse entries removed but testing on the full range of entries, yielded results worse than training on the full dataset.

6.4 Error Analysis

We next analyze the errors made by the transformer model on the full phonetic Germanic dataset. The errors were extracted using the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970), which calculates edit distance but also provides an alignment of the two sequences by backtracking through the dynamic algorithm matrix. During this backtracking, insertions, deletions and replacements are identified.

The most common source of errors by far was the inability for the model to predict verb and noun endings. Verb infinitives in Proto-Germanic most commonly have the endings '-aną' or '-ijaną'. The model was often unable to predict correctly which verb ending was correct, and so a large number of insertions and deletions matched the 'ij' difference between the two endings. The other main source of errors corresponds to noun endings which in the nominative, singular form found in our dataset may take any of the following endings: '-az', '-iz', '-uz', '-ō', '-ô'. The model's inability to correctly predict which class of noun a reconstruction belonged to was reflected in the many insertions, deletions, and replacements resulting from confusion between the endings.

Another conspicuous source of deletions was that of the phoneme /x/, written 'h' in the standard PGmc orthography. This phoneme appears to have been elided or merged with other sounds in most of the Germanic languages and does not appear in modern pronunciation. For example, take the PGmc word '*tanhuz', cognate with English 'tough', Dutch 'taai', and German 'zäh'. While the orthography of English and German preserve the sound, in English the corresponding digraph 'gh' is now pronounced /f/, while the German 'h' is silent. In fact, deletion of the 'h' graph from the orthographic reconstructions are much rarer than for the corresponding /x/ phoneme, showing how archaic pronuncations are preserved in orthography. The problem of phoneme elision is one that linguists also struggle with as it is very difficult to reconstruct a sound that is no longer present in a modern language. This problem is further exacerbated in machine learning methods that lack the domain knowledge of how elided sounds leave vestigial traces in modern pronunciations.

A similar set or errors arose with the phoneme /n/ which was inserted and deleted in many of the predictions. This can be traced to the fact that Proto-Germanic has nasalized vowels unlike any of its daughter languages. These nasalizations were either removed or fully pronounced as /n/ in the daughter languages, making it difficult to predict their occurrences.

Finally, we see how exclusively giving the transformer model phonetic information is a limitation. For most of the compound words included in the dataset (e.g. winedrunken, starblind, mereswine, neighbor) the model gives wildly inaccurate predictions. The model is unable to infer the morphological information that linguists can directly

utilize.

6.5 Pattern Learning

Despite the errors, the transformer model does successfully pick up on many of the historical patterns of phonetic change. A sample of these are shown in Table 5. The Proto-Germanic sounds 'sw', 'b', 'au' and 'eu' have systematically diverged, consistently showing up in the modern languages in their diverged forms. Table 5 shows how the model is able to correctly predict the correct ancestral sound from their modern correspondences. Even for vowels, where some of the sound shifts are less consistent, the model successfully infers the patterns and applies them to its predicted reconstruction. This is the essence of the comparative method for reconstructing proto-forms, which the model has managed to capture.

We can also tell that the model has truly learned these patterns because it applies them too often. The model wrongly predicts the reconstructions '*seuniz' and '*krabjanaą' for the golden Proto-Germanic'*siuniz' and '*krafjaną' as it notices the same sounds in the daughter cognates that would normally correspond to a 'iu' or a 'b' respectively. This is a clear indication that the model has learned phonetic divergence patterns.

Aside from orthographic and phonetic patterns, the model is also able to pick out some morphological patterns. The chief example of this is how the model recognizes how German and Dutch verb infinitives ending in '-en' correspond to Proto-Germanic infinitives ending in '-aną' or '-ijaną'. Whenever a Dutch and a German word ending in '-en' appeared in the input, the model would predict either verb ending with 94.5%accuracy.

6.6 Novel Predictions

Table 6 contains a series of cognates for which no PGmc reconstruction is available. No corresponding PGmc entries for these words exist neither in Wiktionary nor in Kroonen's (2013) dictionary. For each set of the cognates in Table 6 we had the transformer model predict a PGmc reconstruction. Of course, we should not take these predictions as ground truth and a linguist would have to verify their correctness. However, this remains a proof of concept for how machine learning models can aid linguists in reconstructing proto-forms when a full set of cognates is not available. And, in the author's opinion seem quite convincing.

7 Conclusion

We have created the first dataset of Germanic language cognates which we hope will be used to further research into automatic proto-from reconstruction. While initially under-performing when used to train state-of-the-art proto-from reconstruction models, we show that removing entries with a small number of cognates improves the robustness of the dataset as we effectively remove spurious patterns. In the end, we managed to train a model with resulting normalized edit distance better than Kim et al.'s (2023) Sinitic dataset.

Moreover, we show how a transformer model trained on our Germanic dataset produces regular errors arising from lost phonemic information in the daughter languages. Conversely, the same model is able to learn patterns of historical sound shifts to accurately predict certain phonemes in Proto-Germanic words. Lastly, we show models trained on our dataset can be used to produce novel Proto-Germanic reconstructions found nowhere else in the literature.

7.1 Limitations and Further Work

The main limitation of the dataset is sparsity of many of the cognate-sets. As explored in Section 6.3 models trained on a reduced dataset with fuller cognate-sets performed much better. The issue of sparsity can be addressed by expanding the set of languages we include in our dataset beyond the 7 used in this paper. Moreover, by including other ancestral languages like Old English and Old Norse that are closer to Proto-Germanic we should expect a marked improvement in performance. We can additionally make use of other etymological resources beyond Wiktionary to extend the number of entries in the dataset. However, the lack of good APIs for most dictionaries would make collecting information more difficult and time consuming.

Moreover, a key source of error was the lack of morphological and grammatical information in the dataset. Using the architecture of the models discussed, there is no way of accurately predicting the noun and verb endings present in Proto-Germanic from the phonetics and orthography of its descendant languages. An alternate architecture such as that described by List et al. (2022b), allows for additional information, such as the part-of-speech tags, to be included in the input sequences. Including noun gender, or verb irregularity information could give the machine learning model the extra

Danish	Dutch	English	German	Icelandic	Swedish	Predicted PGmc	Gold PGmc
-	**zw**eer	-	**Schw**äher	-	**sv**är-	**sw**erhō	**sw**ehuraz
sværd	**zw**aard	**sw**ord	**Schw**ert	**sv**ërð	**sv**ärd	**sw**erdą	**sw**erdą
-	**zw**inden	-	**Schw**inden	-	**sv**inna	**sw**indaną	**sw**indaną
-	**zw**ingen	**sw**ing	**Schw**ingen	-	**sv**inga	**sw**inganą	**sw**inganą
-	**zw**ijmen	-	-	-	**sv**imma	**sw**īmaną	**sw**īmaną
væ**v**e	we**v**en	wea**v**e	we**b**en	vä**v**a	ve**f**a	wa**b**janą	we**b**aną
na**v**le	na**v**el	na**v**el	na**b**el	na**f**li	na**v**el	na**b**ulaz	na**b**alô
søl**v**	zil**v**er	sil**v**er	sil**b**er	sil**f**ur	sil**v**er	silu**b**raz	silu**b**rą
forgi**v**e	verge**v**en	forgi**v**e	verge**b**en	fyrirge**f**a	förgi**v**a	frage**b**aną	frage**b**aną
kræ**v**e	-	cra**v**e	-	kre**f**ja	krä**v**a	kra**b**janą	kra**f**janą
k**ø**be	k**o**pen	ch**ea**p	k**äu**fen	k**ey**pa	k**ö**pa	k**au**pijaną	k**au**pijaną
-	-	-	-	sm**ey**ja	sm**ö**ja	sm**au**gijaną	sm**au**gijaną
-	t**o**men	t**ee**m	z**äu**men	t**ey**ma	-	t**au**mijaną	t**au**mijaną
-	h**o**nen	h**ea**n	h**ö**hnen	-	-	h**au**nijaną	h**au**nijaną
-	z**o**men	-	s**äu**men	s**ey**ma	-	s**au**mijaną	s**au**mijaną
s**y**de	z**ie**den	s**ee**the	s**ie**den	s**jó**da	s**ju**da	seuþaną	seuþaną
l**y**d	-	-	-	hl**jó**d	l**ju**d	hle**u**þą	hle**u**þą
b**y**de	b**ie**den	b**i**d	b**ie**ten	b**jó**da	b**ju**da	b**eu**daną	b**eu**daną
g**y**de	g**ie**ten	g**u**t	g**ie**ten	g**jó**ta	g**ju**ta	g**eu**taną	g**eu**taną
s**y**n	-	-	-	s**jó**n	s**y**n	s**eu**niz	s**iu**niz

Table 5: The transformer model's predictions of regular sound patterns across the cognates. The model overextends these patterns in the case of '*siuniz' and '*krafjaną'.

Danish	Dutch	English	German	Icelandic	Swedish	Predicted PGmc
-	grillen	grill	grellen	-	gräla	grelaną
krus	kroes	cruse	Krause	krús	krus	krūs
-	laven	lave	laben	-	-	labōną
nar	nar	-	Narr	narri	narr	narzô
snarke	snerken	snark	schnarchen	-	snarka	snarkijaną

Table 6: Novel predictions of Germanic cognates.

information needed to deduce these morphological patterns.

Finally, beyond the scope of this paper, there is an interest in developing automatic cognate identification methods in order to create cognate datasets, especially for less-resourced and extinct languages where a small lexicon without any etymological information may be all we have. In our case, such a system may help identify previously undiscovered cognates to further extend the dataset. Similarly, whenever we do not have a golden reconstruction we have to rely on unsupervised proto-form reconstruction. State-of-the-art unsupervised models still lag behind supervised models (He et al., 2022). Performing a similar analysis an unsupervised model trained on variations of our dataset should give us a better understanding of how we should design our datasets to optimize unsupervised learning.

References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Alexandre Bouchard-Côté, David Hall, Thomas Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National*

Academy of Sciences of the United States of America, 110.

Kalvin Chang, Chenxuan Cui, Youngmin Kim, and David R. Mortensen. 2022. WikiHan: A new comparative dataset for Chinese languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3563–3569, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Liviu Dinu and Alina Maria Ciobanu. 2014. Building a dataset of multilingual cognates for the Romanian lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1038–1043, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andre Wang He, Nicholas Tomlin, and Dan Klein. 2022. Neural unsupervised reconstruction of protolanguage word forms. *ArXiv*, abs/2211.08684.

Young Min Kim, Kalvin Chang, Chenxuan Cui, and David R. Mortensen. 2023. Transformed protoform reconstruction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–38, Toronto, Canada. Association for Computational Linguistics.

Guus J Kroonen. 2013. *Etymological Dictionary of Proto-Germanic*. Brill.

Winfred P. Lehmann. 2014. *A Grammar of Proto-Germanic*. Linguistics Research Center, University of Texas at Austin.

Stephen Pax Leonard. 2011. Relative linguistic homogeneity in a new society: The case of iceland. *Language in Society*, 40(2):169–186.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Johann-Mattis List. 2019. Beyond edit distances: Comparing linguistic reconstruction systems. *Theoretical Linguistics*, 45(3-4):247–258.

Johann-Mattis List, Robert Forkel, Simon Greenhill, Christoph Rzymski, Johannes Englisch, and Russell Gray. 2022a. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9:316.

Johann-Mattis List, Robert Forkel, and Nathan Hill. 2022b. A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 89–96, Dublin, Ireland. Association for Computational Linguistics.

Johann-Mattis List, Ekaterina Vylomova, Robert Forkel, Nathan Hill, and Ryan Cotterell. 2022c. The SIGTYP 2022 shared task on the prediction of cognate reflexes. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 52–62, Seattle, Washington. Association for Computational Linguistics.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473, Online. Association for Computational Linguistics.

D. Gary Miller. 2019. Alphabet and phonology. In *The Oxford Gothic Grammar*. Oxford University Press.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Frans van Coetsem and Herbert L. Kufner. 1972. *Toward a Grammar of Proto-Germanic*. Walter de Gruyter GmBH.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.