The SIGTYP 2022 Shared Task on the Prediction of Cognate Reflexes

Johann-Mattis List[ɯ] **Ekaterina Vylomova**[ɘ] **Robert Forkel**[ɯ]
Nathan W. Hill[ʌ] **Ryan D. Cotterell**[ð]
[ɯ]MPI-EVA Leipzig [ɘ]University of Melbourne [ʌ]University of Dublin [ð]ETH Zürich
mattis_list@eva.mpg.de

Abstract

This study describes the structure and the results of the SIGTYP 2022 shared task on the prediction of cognate reflexes from multilingual wordlists. We asked participants to submit systems that would predict words in individual languages with the help of cognate words from related languages. Training and surprise data were based on standardized multilingual wordlists from several language families. Four teams submitted a total of eight systems, including both neural and non-neural systems, as well as systems adjusted to the task and systems using more general settings. While all systems showed a rather promising performance, reflecting the overwhelming regularity of sound change, the best performance throughout was achieved by a system based on convolutional networks originally designed for image restoration.

1 Introduction

In historical-comparative linguistics, scholars typically assemble words from related languages into *cognate sets*. In contrast to the notion of cognacy in language teaching and synchronic NLP applications, cognate sets are understood as sets of words that share a common origin regardless of their meaning in historical-comparative linguistics and that should not contain borrowed words. The individual members of a cognate set are typically called *cognate reflexes* or simply *reflexes* (Trask, 2000, 278). Cognate reflexes typically show regular sound correspondences. This means that one can define a mapping across the individual phoneme systems of the individual languages. Thus, English *t* typically corresponds to a German *ts* (compare *ten* vs. *zehn*), and English *d* corresponds to German *t* (compare *dove* vs. *Taube*). The mappings often depend on certain contextual conditions and may differ, depending on the position in which they occur in a word. With the help of regular sound correspondences, linguists can often predict fairly well how the cognate counterpart of a word in one language might sound in another language. However, prediction by linguists rarely takes only one language pair into account. The more reflexes a cognate set has in different languages, the easier it is to predict reflexes in individual languages.

1.1 The Reflex Prediction Task

In its simplest form, the data we need for the task of reflex prediction is a table in which each column represents a different language and each row a different cognate set. We also assume that word forms (or "reflexes" of a cognate set) are represented in standardized phonetic transcriptions (such as the International Phonetic Alphabet). Whenever a reflex in a specific language is missing, this reflex can in theory be predicted with the help of the remaining reflexes. As an example, consider Table 2, showing reflexes of cognate sets in German, English, and Dutch. Since the reflex for the BELLY cognate sets is missing in English, we could try and predict it from known correspondences to German and Dutch. The correct prediction would be *bouk*. This form has been still preserved for some time in English in the meaning of "torso", going back to Old English *būk* "belly" (Pfeifer, 1993), although it has nowadays come out of use. When provided with more data of this kind, one can build a model that would be able to predict an English form given a German and a Dutch form, as well as a German form, given a Dutch and an English form, and so on. Note that not all cognate sets in real-life data will have reflexes for all words. Thus, we know about English *bouk* from dialect records, but without dialects or written sources from Middle English, we could only rely on prediction itself in order to guess how the word would sound if it would have been retained.

Since predictions for words that have been completely lost cannot be evaluated directly, we will base our task on the prediction of artificially ex-

Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP (SIGTYP 2022), pages 52 - 62
July 14, 2022 ©2022 Association for Computational Linguistics

Training Data

Dataset	Source	Version	Family	Languages	Words	Cognates
⋆abrahammonpa	Abraham (2005)	v3.0	Tshanglic	8	2063	403
⋆allenbai	Allen (2007)	v4.0	Bai	9	5773	969
⋆backstromnorthernpakistan	Backstrom and Radloff (1992)	v1.0	Sino-Tibetan	7	1426	248
⋆castrosui	Castro and Pan (2015)	v3.0.1	Sui	16	10139	1048
davletshinaztecan	Davletshin (2012)	v1.0	Uto-Aztecan	9	771	118
felekesemitic	Feleke (2021)	v1.0	Afro-Asiatic	19	2583	340
⋆hantganbangime	Hantgan and List (2018)	v1.0	Dogon	16	4405	971
hattorijaponic	Hattori (1973)	v1.0	Japonic	10	1802	278
listsamplesize	List (2014)	v1.0	Indo-European	4	1320	512
mannburmish	Mann (1998)	v1.2	Sino-Tibetan	7	2501	576

Surprise Data

Dataset	Source	Version	Family	Languages	Words	Cognates
bantubvd	Greenhill and Gray (2015)	v4.0	Atlantic-Congo	10	1218	388
beidazihui	Běijīng Dàxué (1962)	v1.1	Sino-Tibetan	19	9750	518
birchallchapacuran	Birchall et al. (2016)	v1.1.0	Chapacuran	10	939	187
bodtkhobwa	Bodt and List (2022)	v3.1.0	Western Kho-Bwa	8	5214	915
⋆bremerberta	Bremer (2016)	v1.1	Berta	4	600	204
⋆deepadungpalaung	Deepadung et al. (2015)	v1.1	Palaung	16	1911	196
hillburmish	Gong and Hill (2020)	v0.2	Sino-Tibetan	9	2202	467
kesslersignificance	Kessler (2001)	v1.0	Indo-European	5	565	212
luangthongkumkaren	Luangthongkum (2019)	v0.2	Sino-Tibetan	8	2363	379
⋆wangbai	Wang and Wang (2004)	v1.0	Sino-Tibetan	10	4356	658

Table 1: Training and surprise data data used in our study. Datasets with identifiers preceded by an asterisk are those in which we automatically searched for cognates. The remaining datasets all provided expert cognates, which we used for the shared task. All datasets are archived with Zenodo, and the supplementary material provides a direct reference to their Zenodo DOI and their GitHub repository URLs.

cluded word forms. Thus, we first take a dataset with cognates in a few related languages, and then artificially delete some of the words in the datasets, using varying proportions. When training a model to predict the missing word forms, we can then compare the predicted words directly with the words we have deleted automatically (List, 2019a).

A special case of the reflex prediction task, *supervised phonological reconstruction*, focuses on the prediction of words in ancestral languages, thus mimicking the process of *phonological reconstruction* as one of the key aspects of the traditional *comparative method* (Weiss, 2015). While we predict reflexes in any language in the generic reflex prediction task, in automated phonological reconstruction we predict one specific reflex of a cognate set, viz. the form in the ancestral language. Apart from the restriction in scope, however, the two tasks do not differ much, and most methods which solve the one task could also be used to solve the other.

Cognate Set	German	English	Dutch
ASH	a ʃ ɛ	æʃ	ɑ s
BITE	b ai s ə n	b ai t	b ɛi t ə
BELLY	b au x	-	b œi k

Table 2: Exemplary cognate reflexes in German, English, and Dutch.

1.2 Background on Reflex Prediction

Quite a few studies on cognate reflex prediction have been published during the past years. Beinborn et al. (2013) uses character-based machine translation approaches to predict cognate candidates in a bilingual setting. Bodt and List (2022) use a method for cognate reflex prediction originally tested by List (2019a) to predict cognate reflexes in so far unobserved data, which was later verified in fieldwork. The method by List (2019a) uses automatically identified *sound correspondence patterns* and phonetic alignment analyses in order to predict for a given set of cognate words how reflexes in languages missing in the cognate set would sound. Meloni et al. (2021) make use of an encoder-decoder model in order to reconstruct Latin words from cognate sets in Romance languages. Fourrier et al. (2021) model cognate reflex prediction as a low-resource machine translation task, building several translation models for Romance languages and using these to evaluate word prediction accuracy. Dekker and Zuidema (2021) use recurrent neural networks for cognate reflex prediction and illustrate how word prediction can be used to solve additional tasks in computational historical linguistics, such as phylogenetic reconstruction or sound correspondence detection.

List et al. (2022a) build on the framework for sound correspondence pattern detection by List (2019a) in order to propose a new framework for supervised phonological reconstruction and cognate reflex prediction which they expand by *enriching* phonetic alignment analyses in such a way that contextual information can be taken account.

1.3 Difficulties of Reflex Prediction

For traditional as well as modern approaches to reflex prediction, there are a couple of challenges that algorithmic solutions need to account for. The first challenge consists in the prediction of sounds which have no corresponding counterpart in the source languages from which one predicts a word in the target language. As an example, consider Dutch *tand* [t ɑ n d] "tooth" and English *tooth* [t ʊː θ]. It is easy to see that the [t] in Dutch corresponds to a [t] in English, such as [aː] corresponds to [ʊː] and [θ] corresponds to [d]. However, the [n] in Dutch has no counterpart in English, since English [n] was lost when followed by a [θ]. Since there is no one-to-one sound match between the sound in English and the sound in Dutch, the prediction has to be based on the *conditioning context*, which is notoriously difficult to handle in computational approaches.

A further difficulty consists in the *sparsity* and the *patchiness* of the data. Data are *sparse* with respect to the number of cognate sets which we can use to train computers or humans. Even for well-established language groups, etymological dictionaries, which list more than 1000 reconstructed items are quite rare. Apart from being *sparse*, data are also *patchy*. Only a very small amount of the proto-forms listed in etymological dictionaries is reflected in the majority of the branches, and an even smaller amount has survived without notable irregularities in the sound changes or the morphology of the word forms. Thus, even if one works with datasets consisting of large numbers of related words, there will always be situations in which important reflexes are missing and at times only one witness may be left that we can use for the prediction of the cognate reflex in question.

2 Materials and Methods

2.1 Materials

Data for the shared task were taken from the Lexibank repository, which offers wordlists from 100 standardized datasets (List et al. 2022a, https://github.com/lexibank/lexibank-analysed). In this repository, a large collection of datasets with cognate sets provided by experts and phonetic transcriptions added by the Lexibank team are provided. An even larger number of datasets has only standardized phonetic transcriptions but no cognate judgments. Since cognate detection methods work well by now, we can determine the cognates specifically for shallower language families with quite some confidence; this enabled us to assemble a larger amount of datasets from different language families and either use cognate sets provided by experts or inferring cognates ourselves, using state-of-the-art methods for automated cognate detection implemented in the LingPy software library (List and Forkel, 2021).

For each the training and the surprise phase, 10 datasets were selected. Following the Lexibank workflow for the curation of lexical wordlists, all datasets were curated on GitHub and additionally archived with Zenodo. Standardization of the data included mapping the language names to Glottolog (Hammarström et al., 2021), linking the concept elicitation glosses to the Concepticon reference catalog (https://concepticon.clld.org, List et al. 2022c), and adding standardized phonetic transcriptions, following the B(road)IPA system of the Cross-Linguistic Transcription Systems reference catalog (https://clts.clld.org, Anderson et al. 2018), with the help of orthography profiles (Moran and Cysouw, 2018). Since only a smaller number of the datasets came along with suitable cognate judgments needed for the cognate reflex prediction task, cognates were automatically inferred with standard settings, using a variant of the LexStat algorithm for automatic cognate detection (List, 2012a) that searches for partial rather than full-word cognates (List et al., 2016). Searching for partial cognates is justified, since both the identification of regular sound correspondences and the prediction of cognate reflexes can only be carried out on material that is entirely cognate (Schweikhard and List, 2020). Since full-word cognates may often contain non-cognate material, the prediction of full cognates would unnecessarily exacerbate the reflex prediction task, adding a random component that cannot be handled algorithmically in a principled way. In all cases, we excluded all singleton cognate sets (cognate sets that occur only in one language), since these cannot be used in our prediction experiments. Table 1 lists all datasets for the test and training phase along with some basic statistics.

The datasets were used as the basis for the data used for test and training during our shared task. For this purpose, each dataset was split into five training and test partitions in which the data retained for testing was varied, starting from a proportion of 10% retained for testing (proportion 0.1), followed by 20% (proportion 0.2), 30% (proportion 0.3), 40% (proportion 0.4), and finally 50% (proportion 0.5). The training data was not modified further and used as primary input for the training phase of all systems. The test data, however, was artificially constructed from the test partition. We first iterated over all cognate sets and then created individual test sets from each cognate set iterating over all words in a cognate set and deleting each word in a row. For a cognate set of *n* words, this would result in *n* test cases, in which each word in each language would have to be predicted one time.

2.2 Methods

2.2.1 Evaluation

Among the most commonly used evaluation measures for the word prediction task is the edit distance, which computes the number of operations needed in order to convert the predicted word into the attested word (Levenshtein, 1965). In its primary form, the edit distance is an integer. In order to normalize the measure, correcting for a bias resulting from the length of the compared strings, scholars have proposed to divide the distance by the length of the largest string (Holman et al., 2008), or by the mean length of both strings being compared (Nerbonne et al., 1999). A further possibility closer to notions of distance in bioinformatics, which we used in our shared task, is to divide the edit distance by the length of the alignment of both strings. The normalized edit distance then corresponds to the normalized Hamming distance between two aligned sequences (Hamming, 1950), or – when subtracting from 1 – to the notion of *percentage identity* in evolutionary biology (Raghava and Barton, 2006). It is, however, important to note that actual differences in these normalization procedures are usually small.

The edit distance, both normalized and unnormalized, has been employed in many word prediction and phonological reconstruction experiments as the basic evaluation measure for the prediction accuracy (Meloni et al., 2021; Bouchard-Côté et al., 2013). Its clearest shortcoming lies in the fact that it only accounts for *surface* differences between prediction and attested words (also called 'phenotypic differences' by Lass 1997), while structural aspects (called 'genotypic differences' by Lass 1997) are ignored. Thus, if a method mistakenly maps a certain sound *x* to a certain sound *y* in all cases in which the *x* occurs, the edit distance will treat each occurrence of the error independently and may therefore provide drastically lowered results. It would, therefore, be good to account for the relative *regularity* of the co-occurrence of *x* and *y*. List (2019b) proposes to compute B-Cubed F-scores (Amigó et al., 2009) from the aligned predicted and attested words. B-cubed F-scores only check for the regularity of occurrences. This results in scores of 1 (indicating complete identity) for sequence pairs like `abbc` compared with `1223`. Indeed, both sequences are structurally completely identical since a simple mapping between the symbols in both sequences can convert one string into the other and vice versa. If a method has systematic errors but otherwise does a good job in prediction, B-Cubed F-Scores penalize results less strongly than edit distance. As a final evaluation score, we followed Fourrier et al. (2021) in providing BLEU scores (Papineni et al., 2002). These scores are usually used to investigate how well an automated translation corresponds to the translated target test. BLEU scores and B-Cubed F-Scores range from 0 to 1, with 1 indicating perfect agreement, the normalized edit distance ranges between 1 (maximal difference) and 0 (string identity).

2.2.2 Baselines

Our baselines were taken from the reflex prediction framework by List et al. (2022b). This framework consists of four major stages. In stage (1), cognate sets are aligned with the help of standard methods for multiple phonetic alignment analyses (List, 2012b). In stage (2), alignments are *trimmed* by merging all columns in the alignment in which the attested languages all show a gap with their preceding column. As a result, a word like Latin *cenāre* [k eː n aː r ɛ] would be rendered as [k eː n aː r.ɛ], when being aligned with Spanish *cenar* [θ e n a ɾ], since the final [ɛ] in Latin corresponds to a gap in Spanish and could therefore not be predicted (see List et al. 2022b for details on this procedure). In stage (3), alignments are *enriched* by coding for potentially conditioning context, which is added to the alignments in the form of additional rows. In stage (4), the individual alignment columns are converted to a matrix from which a classifier can be trained. During prediction, cognate sets fed to the algorithm are again being aligned and enriched, but the trimming procedure is not needed, since it only relates to the target language that one wants to predict.

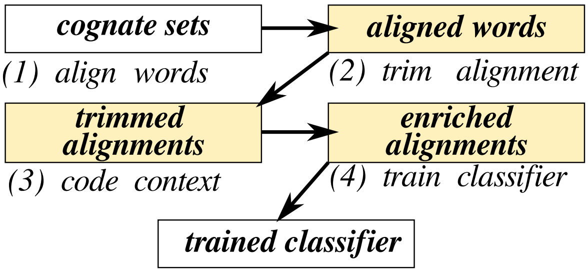

Figure 1: Major steps of the reflex prediction framework underlying the baseline.

Based on this general framework, we created two baselines, one primary baseline that uses the correspondence pattern recognition (**CORPAR**) method by List (2019a) as a classifier, and one extended baseline which predicts words with the help of a support vector machine (**SVM**, see List et al. 2022b for details on both systems). From previous studies on supervised phonological reconstruction we know that the **SVM** variant of the framework outperforms the **CORPAR** classifier clearly, although differences are not extremely high (ibid.).

2.2.3 Implementation

We created specific software package that allows to (1) automatically download the data in the particular versions of the individual CLDF datasets which we used, (2) create the test and training data, (3) apply the baseline methods to the data, and (4) carry out the evaluation. The software package is written in Python and can be accessed both using the commandline and from within Python scripts. It is curated on GitHub and archived with Zenodo (see Section Supplementary Material for details). Different versions were created in order to first release the training data (version 1.1), followed by the release of the surprise data (version 1.2), and finally followed by the release of the official results of the evaluation (version 1.4, providing extended evaluations in contrast to the version 1.3 planned earlier).

Major dependencies of the software package are LingPy (List and Forkel, 2021), used for the computation of the edit distance and of phonetic alignments, Lingrex (List and Forkel, 2022), providing access to the baseline method for cognate reflex prediction, Scikit-learn (Pedregosa et al., 2011), providing access to support vector machines, and Matplotlib (Hunter, 2007), used for plotting.

3 Systems

Four teams submitted their systems for our shared task. Since these systems are described in individual papers (Kirov et al., 2022; Jäger, 2022; Tresoldi, 2022; Celano, 2022), we will only briefly present their main features here.

Team CrossLingference, represented by Gerhard Jäger (University Tübingen), provided a workflow Jäger 2022, implemented in the **JULIA** programming language, that makes specifically use of Bayesian phylogenetic inference. In contrast to the remaining systems submitted to our shared task, Jäger's approach takes phylogenetic information into account, extending an earlier workflow for phonological reconstruction (Jäger, 2019).

Team Mockingbird, represented by Christo Kirov, Richard Sproat, and Alexander Gutkin (Google Research), provided two models for the prediction of cognate reflexes. The first model, the **NEIGHBOR TRANSFORMER MODEL**, was originally designed to find problems in the readings of Japanese place names spelled in kanji (Jones et al., 2022), and is based on the popular transformer architecture (Vaswani et al., 2017), which was specifically adjusted for the task. Since the training data would be too small for the transformer model, the authors augmented it with new instances generated by randomly sampling subsets of a corresponding cognate set. In addition to that, they also enriched each set with synthetic instances using n-gram language modelling. The second model, the **IMAGE INPAINTING MODEL**, compares the cognate reflex prediction task to the task of restoring corrupted parts of a 2D image, in which dimensions correspond to languages and cognate phonemic representations. The restoration is achieved with the help of convolutional neural networks (Liu et al., 2018). For this model, no data augmentation steps were undertaken. The authors provide four model configurations of the neighbor model, with the first three (**N1-A**, **N1-B**, and **N1-C**) differing in the number of training steps and not being publicly released, while the last one (**N2**), which was only applied to the 0.1 proportion of the data, being publicly released. For the image inpainting model, one configuration was provided (**I1**).

Team Leipzig, represented by Giuseppe G. A. Celano (University Leipzig), provided a TRANSFORMER-based architecture with character and position embeddings for the prediction of cognate reflexes (Vaswani et al., 2017), in which language information was one-hot encoded and the model was trained on individual reflex pairs on each language independently. In order to predict a word from several reflexes in different languages, the system first predicts individual target tensors of probabilities for each attested reflex and then averages them to produce the prediction.

Team CEoT, represented by Tiago Tresoldi (Uppsala University), provided a workflow that predicts cognate reflexes based on phonetic alignments (Tresoldi, 2022), which is quite similar to the extended baselines of our shared task (List et al., 2022b). In contrast to our baseline approaches, their system **EXTALIGN-RF** skips the trimming procedure (stage 2), varies the techniques for alignment enrichment by taking preceding and following context into account (stage 3), and uses a random forests classifier rather than a support vector machine.

While all teams tried hard to provide results for all of their systems, some results could not be computed in time, to be included in the shared task. All teams were asked to share their data in such a way that users can easily replicate the results and also apply their methods to new data. Unfortunately, there was no time for the team organizing the shared task to individually check all systems with respect to replicability and transparency. The team checked, however, that all systems were properly archived with repositories offering long-term storage of data, such as Zenodo, and we communicated the importance of replication with all authors.

Proportion in Test: 0.1

System	ED	NED	B-Cubes	BLEU
Baseline	1.2095	0.3119	0.7231	0.5716
Baseline-SVM	**1.0189**	**0.2625**	**0.7626**	0.6387
CEoT-Extalign-RF	1.0377	0.2763	0.7475	0.6243
CrossLingference-Julia	1.4804	0.3929	0.7251	0.4793
Leipzig-Transformer	1.3901	0.3687	0.6489	0.5114
Mockingbird-I1	0.9201	0.2431	0.7673	0.6633
Mockingbird-N1-A	1.0223	0.2568	0.7604	**0.6479**
Mockingbird-N1-B	1.0437	0.2625	0.7572	0.6398
Mockingbird-N1-C	1.1263	0.2867	0.7302	0.6115
Mockingbird-N2	1.2095	0.3135	0.7054	0.5744

Proportion in Test: 0.2

System	ED	NED	B-Cubes	BLEU
Baseline	1.3253	0.3361	0.6680	0.5412
Baseline-SVM	1.1723	0.2928	**0.7067**	0.5985
CEoT-Extalign-RF	1.2208	0.3175	0.6798	0.5709
CrossLingference-Julia	1.4954	0.3912	0.6882	0.4760
Leipzig-Transformer	1.5787	0.4046	0.5683	0.4646
Mockingbird-I1	1.0413	0.2648	0.7120	0.6326
Mockingbird-N1-A	**1.1512**	**0.2825**	0.7011	**0.6138**
Mockingbird-N1-B	1.1726	0.2901	0.6910	0.6054
Mockingbird-N1-C	1.2196	0.3051	0.6669	0.5841

Proportion in Test: 0.3

System	ED	NED	B-Cubes	BLEU
Baseline	1.4354	0.3556	0.6372	0.5195
Baseline-SVM	1.3713	0.3310	**0.6565**	0.5554
CEoT-Extalign-RF	1.4038	0.3525	0.6331	0.5286
CrossLingference-Julia	1.6116	0.4130	0.6508	0.4503
Leipzig-Transformer	1.7746	0.4467	0.5129	0.4207
Mockingbird-I1	1.1762	0.2899	0.6717	0.6059
Mockingbird-N1-A	**1.2565**	**0.3119**	0.6557	0.5779
Mockingbird-N1-B	1.2712	0.3103	0.6531	**0.5792**
Mockingbird-N1-C	1.3009	0.3215	0.6343	0.5636

Proportion in Test: 0.4

System	ED	NED	B-Cubes	BLEU
Baseline	1.6821	0.4011	0.6001	0.4717
Baseline-SVM	1.6159	0.3891	0.5990	0.4903
CEoT-Extalign-RF	1.5695	0.3960	0.5805	0.4773
CrossLingference-Julia	1.6059	0.4112	**0.6411**	0.4473
Leipzig-Transformer	1.9221	0.4800	0.4736	0.3893
Mockingbird-I1	1.2725	0.3162	0.6428	0.5724
Mockingbird-N1-A	**1.4542**	**0.3521**	0.6294	0.5293
Mockingbird-N1-B	1.3618	0.3349	0.6212	**0.5466**
Mockingbird-N1-C	1.4353	0.3547	0.5999	0.5228

Proportion in Test: 0.5

System	ED	NED	B-Cubes	BLEU
Baseline	1.8889	0.4445	0.5617	0.4265
Baseline-SVM	1.9330	0.4619	0.5371	0.4204
CEoT-Extalign-RF	1.8434	0.4576	0.5194	0.4128
CrossLingference-Julia	1.6794	0.4274	0.6193	0.4296
Leipzig-Transformer	2.1036	0.5257	0.4306	0.3438
Mockingbird-I1	1.4170	0.3518	**0.6050**	0.5337
Mockingbird-N1-A	1.5527	0.3800	0.5959	0.4934
Mockingbird-N1-B	**1.5066**	**0.3734**	0.5864	**0.4989**
Mockingbird-N1-C	1.5818	0.3950	0.5610	0.4749

Table 3: Results for the varying proportions and our four evaluation measures, edit distance (ED), normalized edit distance (NED), B-Cubed F-scores (B-Cubes) and BLEU Scores (BLEU) on the surprise data. Cells shaded in gray highlight the best score obtained for a given proportion, bold font marks the second best score.

4 Results

Given that we measure system performance with four evaluation measures (edit distance, normalized edit distance, B-Cubed F-Scores, and BLEU scores adjusted for word prediction), one might expect that systems perform differently with respect to different evaluation measures. As can be seen from the results in Table 3, however, the results are rather clearly favoring the system **I1** by the team Mockingbird as the winner in almost all proportions. The only case where the Mockingbird **I1** system does not show the best performance is the test with 50% of the words being retained for test-

System	Rank	NED	B-Cubes	BLEU	Aggregated
Mockingbird-I1	1	1	1.2	1	1.1 ± 0.3
Mockingbird-N1-A	2	2.6	3	2.6	2.7 ± 0.4
Mockingbird-N1-B	3	2.4	4	2.4	2.9 ± 0.9
Baseline-SVM	4	5.2	4	5	4.7 ± 1.9
Mockingbird-N1-C	5	4.6	6.6	4.6	5.3 ± 1.3
CEoT-Extalign-RF	6	6	7	6.2	6.4 ± 1.1
CrossLingference-Julia	7	7.6	4	7.6	6.4 ± 2.5
Baseline	8	6.8	6.2	6.8	6.6 ± 0.8
Leipzig-Transformer	9	8.8	9	8.8	8.9 ± 0.4

Table 4: Overview of the average ranks of all nine systems for the different dataset proportions along with aggregated ranks.

ing (proportion 0.5), where the **Julia** system by the CrossLingference team shows the best performance with respect to the B-Cubed F-Scores. Since B-Cubed F-Scores emphasize the systematicity of the prediction quality rather than the accuracy in individual cases, we can see that the **Julia** system copes better with systematic aspects of the word prediction tasks in those cases, where the data for the training of the system is limited. That the different scoring systems show at least some degree of independence can also be seen in Figure 2, which shows results for the 10% partition, where the **Julia** system performs worst with respect to edit distances and BLEU scores, while showing a better performance than **N2**, **Transformer**, and the baseline in B-Cubed F-Scores.

While the **SVM** baseline shows a surprisingly good performance on the lowest proportion of data excluded and retained for testing (proportion 0.1), it looses ground with more data excluded for testing. Here, the **N1-A** and **N1-B** systems, again from the Mockingbird team, show the best performance.

Table 4 provides the aggregated ranks for the normalized edit distance, the B-Cubed F-Scores, and the BLEU scores for all systems obtained for all splits of the data. The classical edit distance was excluded in this overview, since it correlates highly with the normalized edit distance and would therefore artificially increase the overall ranks of systems performing well in this regard. Furthermore, the **N2** system by the Mockingbird team was excluded in this analysis, since results could only be provided for the smallest proportion of words retained for testing (proportion 0.1). For each of the five splits of the data and for each of the methods, we ranked the systems according to their performance and later calculated the average of all ranks for each system on each of the three evaluation methods. The aggregated ranks, in which all three evaluation measures are ranked equally, allow us to rank the overall performance of all systems. It shows the overall superiority of the **I1** system of the Mockingbird team, followed by the teams' **N1-A** and **N1-B** methods. The **SVM** baseline and the **N1-C** method by team Mockingbird follow on places four and five. At the end of these ranks are the **Extalign-RF** system by team CeOT, the **Julia** system by Team CrossLingference, followed by the simple baseline and the **Transformer** approach of team Leipzig.

Overall, all systems do quite a good job at recovering unknown words from their cognate sets, specifically in those cases, where only a small part of the test data was retained for the evaluation process. Judging from our practical experience and independently published results on word prediction experiments (List et al., 2022b; Bodt and List, 2022), B-Cubed F-Scores higher than 0.7 and average edit distances of about 1 provide a good starting point for computer-assisted approaches and can already provide active help in various practical annotation tasks in historical linguistics. Thus, scholars working on the reconstruction of certain language families could use predicted proto-forms and later manually correct them, or field workers could use automatically predicted words when trying to elicit specific lexical items to search for cognate words that might have shifted their meanings.

5 Discussion

It was one of the crucial insights made by historical linguists in the early 19th century (Grimm, 1822; Rask, 1818), that sound change proceeds in a surprisingly regular, systematic manner, affecting all sounds in the lexicon of a language that recur in similar phonotactic positions. Without the systematicity and regularity of sound change, it would not be possible to predict the pronunciation of words in one language based on the pronunciation of cognate words in related languages. While it has been known for a long time to linguists that these kinds of predictions can be made on the basis of historical language comparison, the task of cognate reflex prediction has only recently attracted the attention of scholars working in the field of Natural Language Processing and computational linguistics.

With our shared task on cognate reflex prediction, we hoped to achieve two major goals. On the one hand, we wanted to highlight the importance of classical scholarship for computational applications in historical linguistics and linguistic

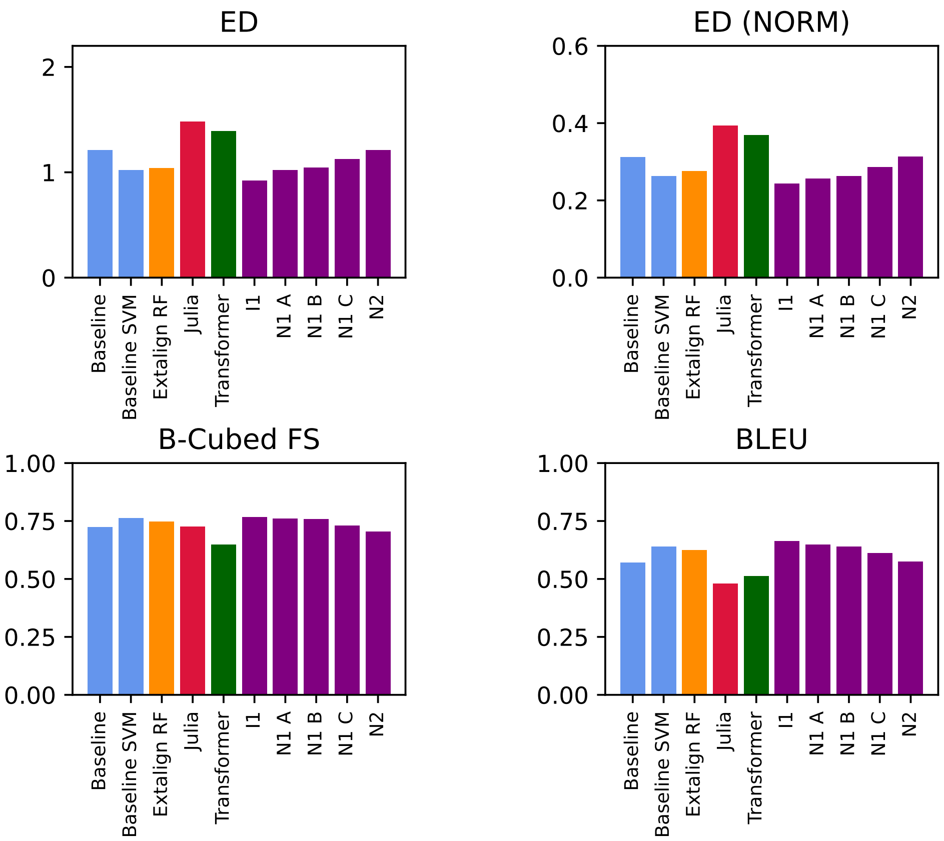

Figure 2: Results for the surprise dataset of the 0.1 proportion, with 10% of data retained for testing.

typology, showing that quite a few problems which are up to today exclusively solved manually might profit from computational treatment. On the other hand, we wanted to trigger the interest of scholars with diverse backgrounds in this task, assembling teams that address the problem with different strategies that might inspire each other and help to lead to largely improved methods in the future.

With the four teams that participated, we have seen an interesting and diverse assembly of systems that all deal with the cognate reflex prediction task. While two teams made use of state-of-the-art machine learning methods based on neural networks (team Mockingbird and team Leipzig), two teams represented systems based on workflows using classical approaches in the emerging discipline of computational historical linguistics (team CrossLingference and team CEoT), using phonetic alignments, and – in the case of team CrossLingference – even Bayesian methods for phylogenetic reconstruction. From the overall performance of the systems in our shared task, we can see that some of the neural approaches outperform the more targeted solutions. Given differences in the performance with respect to the evaluation methods, which highlight different aspects of prediction accuracy, however, we could also see that targeted methods like the Julia method by CrossLingference or the extended Baseline come very close to the best neural systems, and even outperform them at times.

Acknowledgements

This study was partially supported by the Max Planck Society Research Grant "Beyond CALC: Computer-Assisted Approaches to Human Prehistory, Linguistic Typology, and Human Cognition (CALC^3)", awarded to Johann-Mattis List (2022-2024). We thank Clémentine Fourrier for comments and suggestions on the evaluation procedure, and we thank all participants for their feedback during the task.

Supplementary Material

Data and code for the shared task along with results for all systems are curated GitHub (`https://github.com/sigtyp/ST2022`, Version 1.4) and have been archived with Zenodo (`https://doi.org/10.5281/zenodo.6586772`).

References

Binny et al Abraham. 2005. A sociolinguistic research among selected groups in Western Arunachal Pradesh highlighting Monpa. Unpublished manuscript.

Bryan Allen. 2007. *Bai Dialect Survey*. SIL International, Dallas.

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.

Cormac Anderson, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A Cross-Linguistic Database of Phonetic Transcription Systems. *Yearbook of the Poznań Linguistic Meeting*, 4(1):21–53.

Peter C. Backstrom and Carla F. Radloff. 1992. *Languages of Northern Areas*, volume 2 of *Sociolinguistic Survey of Northern Pakistan*. National Institute of Pakistan Studies, Islamabad.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate production using character-based machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 883–891.

Joshua Birchall, Michael Dunn, and Simon J. Greenhill. 2016. A Combined Comparative and Phylogenetic Analysis of the Chapacuran Language Family. *International Journal of American Linguistics*, 82(3):255–284.

Timotheus Adrianus Bodt and Johann-Mattis List. 2022. Reflex prediction. A case study of Western Kho-Bwa. *Diachronica*, 39(1):1–38.

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 of the United States of America*, 110(11):4224–4229.

Nate D. Bremer. 2016. *A sociolinguistic survey of six Berta speech varieties in Ethiopia*. SIL International, Addis Ababa.

Beijing University Běijīng Dàxué. 1962. *Hànyǔ fāngyīn zìhuì 汉语方音字汇 [Chinese dialect character pronunciation list]*. Wénzì Gǎigé, Běijīng.

Andy Castro and Xingwen Pan, editors. 2015. *Sui dialect research*. SIL International, Guizhou.

Giuseppe G. A. Celano. 2022. A Transformer architecture for the prediction of cognate reflexes. In *The Fourth Workshop on Computational Typology and Multilingual NLP*, Online. Association for Computational Linguistics.

Albert Davletshin. 2012. Proto-uto-aztecans on their way to the proto-aztecan homeland: linguistic evidence. *Journal of Language Relationship*, 1(8):75–92.

Sujaritlak Deepadung, Supakit Buakaw, and Ampika Rattanapitak. 2015. A lexical comparison of the Palaung dialects spoken in China, Myanmar, and Thailand. *Mon-Khmer Studies*, 44:19–38.

Peter Dekker and Willem Zuidema. 2021. Word prediction in computational historical linguistics. *Journal of Language Modelling*, 8(2):295–336.

Tekabe Legesse Feleke. 2021. Ethiosemitic languages: Classifications and classification determinants. *Ampersand*, page 100074.

Clémentine Fourrier, Rachel Bawden, and Benoît Sagot. 2021. Can cognate prediction be modelled as a low-resource machine translation task? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 847–861, Online. Association for Computational Linguistics.

Xun Gong and Nathan Hill. 2020. *Materials for an Etymological Dictionary of Burmish*. Zenodo, Geneva.

Simon J Greenhill and Russell D Gray. 2015. Bantu Basic Vocabulary Database.

Jacob Grimm. 1822. *Deutsche Grammatik*, 2 edition, volume 1. Dieterichsche Buchhandlung, Göttingen.

Harald Hammarström, Martin Haspelmath, Robert Forkel, and Sebastiaon Bank. 2021. *Glottolog [Dataset, Version 4.5]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Richard W. Hamming. 1950. Error detection and error detection codes. *Bell System Technical Journal*, 29(2):147–160.

Abbie Hantgan and Johann-Mattis List. 2018. Bangime: Secret language, language isolate, or language island?

Shirō Hattori. 1973. Japanese dialects. In Henry M. Hoenigswald and Robert H. Langacre, editors, *Diachronic, areal and typological linguistics*, number 11 in Current Trends in Linguistics, pages 368–400. Mouton, The Hague and Paris.

Eric W. Holman, Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Advances in automated language classification. In Antti Arppe, Kaius Sinnemäki, and Urpu Nikann, editors, *Quantitative Investigations in Theoretical Linguistics*, pages 40–43. University of Helsinki, Helsinki.

John D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

Llion Jones, Richard Sproat, and Haruko Ishikawa. 2022. Helpful neighbors: Leveraging geographic neighbors to aid in placename pronunciation. In preparation.

Gerhard Jäger. 2019. Computational historical linguistics. *Theoretical Linguistics*, 45(3-4):151–182.

Gerhard Jäger. 2022. Bayesian phylogenetic cognate prediction. In *The Fourth Workshop on Computational Typology and Multilingual NLP*, Online. Association for Computational Linguistics.

Brett Kessler. 2001. *The significance of word lists*. CSLI Publications, Stanford.

Christo Kirov, Richard Sproat, and Alexander Gutkin. 2022. Mockingbird at the SIGTYP 2022 Shared Task: Two types of models for the prediction of cognate reflexes. In *The Fourth Workshop on Computational Typology and Multilingual NLP*, Online. Association for Computational Linguistics.

Roger Lass. 1997. *Historical linguistics and language change*. Cambridge University Press, Cambridge.

Vladimir. I. Levenshtein. 1965. Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov [binary codes with correction of deletions, insertions and replacements]. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Johann-Mattis List. 2012a. LexStat. Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, pages 117–125, Stroudsburg.

Johann-Mattis List. 2012b. SCA: Phonetic alignment based on sound classes. In Marija Slavkovik and Dan Lassiter, editors, *New directions in logic, language, and computation*, pages 32–51. Springer, Berlin and Heidelberg.

Johann-Mattis List. 2014. Investigating the impact of sample size on cognate detection. *Journal of Language Relationship*, 11:91–101.

Johann-Mattis List. 2019a. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics*, 45(1):137–161.

Johann-Mattis List. 2019b. Beyond Edit Distances: Comparing linguistic reconstruction systems. *Theoretical Linguistics*, 45(3-4):1–10.

Johann-Mattis List and Robert Forkel. 2021. *LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.9]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List and Robert Forkel. 2022. *LingRex: Linguistic reconstruction with LingPy [Software Library, Version 1.2]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022a. Lexibank, A public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, pages 1–31.

Johann-Mattis List, Nathan W. Hill, and Robert Forkel. 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*, Dublin [Online]. Association for Computational Linguistics.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*, pages 599–605, Berlin. Association of Computational Linguistics.

Johann-Mattis List, Annika Tjuka, Christoph Rzymski, Simon J. Greenhill, Nathanael E. Schweikhard, and Robert Forkel. 2022c. *Concepticon. A resource for the linking of concept lists [Dataset, Version 2.6.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the 15th European Conference on Computer Vision (ECCV 2018)*, pages 89–105, Munich, Germany. Springer International Publishing. Preprint.

Theraphan Luangthongkum. 2019. A view on Proto-Karen phonology and lexicon. *Journal of the Southeast Asian Linguistics Society*, 12(1):i–lii.

Noel Walter Mann. 1998. *A phonological reconstruction of Proto Northern Burmic*. Phd, The University of Texas, Arlington.

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.

Steven Moran and Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press, Berlin.

John Nerbonne, Wilbert Heeringa, and Peter Kleiweg. 1999. Edit distance and dialect proximity. In David Sankoff and Joseph. B. Kruskal, editors, *Time warps, string edits, and macromolecules. The theory and practice of sequence comparison*, reprint edition, pages V–XV. CSLI Publications, Stanford.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Wolfgang Pfeifer. 1993. *Etymologisches Wörterbuch des Deutschen*, 2 edition. Akademie, Berlin.

G. P. S. Raghava and Geoffrey J. Barton. 2006. Quantification of the variation in percentage identity for protein sequence alignments. *BMC Bioinformatics*, 7(415).

Rasmus K. Rask. 1818. *Undersögelse om det gamle Nordiske eller Islandske sprogs oprindelse*. Gyldendalske Boghandlings Forlag, Copenhagen.

Nathanael E. Schweikhard and Johann-Mattis List. 2020. Developing an annotation framework for word formation processes in comparative linguistics. *SKASE Journal of Theoretical Linguistics*, 17(1):2–26.

Robert L. Trask. 2000. *The dictionary of historical and comparative linguistics*. Edinburgh University Press, Edinburgh.

Tiago Tresoldi. 2022. Approaching reflex predictions as a classification problem from extended phonological alignments. In *The Fourth Workshop on Computational Typology and Multilingual NLP*, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008, Long Beach, CA. Curran Associates Inc.

Feng Wang and William S.-Y. Wang. 2004. Basic words and language evolution. *Language and Linguistics*, 5(3):643–662.

Michael Weiss. 2015. The comparative method. In Claire Bowern and Nicholas Evans, editors, *The Routledge handbook of historical linguistics*, pages 127–145. Routledge, New York.