Evolutionary Insights into Language Change: Cognate Transformer and Phylogenetic Testing in Historical Linguistics

A thesis submitted

in Partial Fulfillment of the Requirements

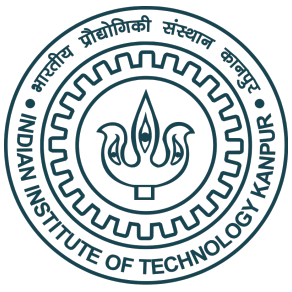
for the Degree of

Doctor of Philosophy

by

Akavarapu V.S.D.S. Mahesh

19111265



to the

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

June, 2024

CERTIFICATE

It is certified that the work contained in the thesis titled **Evolutionary Insights into Language Change: Cognate Transformer and Phylogenetic Testing in Historical Linguistics**, by **Akavarapu V.S.D.S. Mahesh**, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Prof. Arnab Bhattacharya

Department of Computer Science & Engineering

IIT Kanpur

June, 2024

DECLARATION

This is to certify that the thesis titled **Evolutionary Insights into Language Change: Cognate Transformer and Phylogenetic Testing in Historical Linguistics** has been authored by me. It presents the research conducted by me under the supervision of **Prof. Arnab Bhattacharya**. To the best of my knowledge, it is an original work, both in terms of research content and narrative, and has not been submitted elsewhere, in part or in full, for a degree. Further, due credit has been attributed to the relevant state-of-the-art and collaborations (if any) with appropriate citations and acknowledgements, in line with established norms and practices.

Name: Akavarapu V.S.D.S. Mahesh

Programme: PhD

Department: Computer Science & Engineering

Indian Institute of Technology Kanpur

Kanpur 208016

ABSTRACT

Name of student: **Akavarapu V.S.D.S. Mahesh** Roll no: **19111265**

Degree for which submitted: **Doctor of Philosophy**

Department: **Computer Science & Engineering**

Thesis title: **Evolutionary Insights into Language Change: Cognate Transformer and Phylogenetic Testing in Historical Linguistics**

Name of Thesis Supervisor: **Prof. Arnab Bhattacharya**

Month and year of thesis submission: **June, 2024**

Several central problems of historical linguistics, the study of language change, involve strenuous manual labor whether in the case of identification of cognates, reconstruction of proto-languages, or that of classification of languages. Computational historical linguistics, thus, aims to assist the linguistics through automation of these tasks. We introduce novel methods inspired by computational biology that achieve state-of-the-art results on several tasks discussed as follows.

Firstly, we introduce the Cognate Transformer based on the MSA Transformer, a protein language model, to the problems of automated phonological reconstruction and associated cognate reflex prediction. Phonological reconstruction involves predicting the proto-word of an ancestral language from the observed cognate words of daughter languages. In cognate reflex prediction, on the other hand, a reflex word in a daughter language is predicted based on cognate words from other daughter languages. We show that our model outperforms the existing models on both tasks, especially when it is pre-trained on masked word prediction task.

Secondly, we adapt the Cognate Transformer to the task of automated cognate detection by incorporating some modules from Alphafold2, a protein fold predictor. To utilize the labeled information to the fullest, we advocate for a supervised approach where beyond

a certain amount of supervision, the method outperforms existing methods. We also demonstrate that accepting multiple sequence alignments as input and having an end-to-end architecture with a link prediction head saves a significant amount of computation time.

Thirdly, inspired by molecular phylogenetics, we propose a likelihood ratio test to determine if given languages are related based on the proportion of phonetically conserved sites. Lexical resemblances among a group of languages indicate that the languages could be genetically related, i.e., they could have descended from a common ancestral language. However, such resemblances can arise by chance and, hence, need not always imply an underlying genetic relationship. Hence, a genetic relationship should be demonstrated through a significance test. We show that existing multilateral permutation-based tests are prone to yield false positives while our approach overcomes this problem. We demonstrate that the test supports the existence of sub-groupings of two macro-families, namely Mayan-Mixe-Zoque in the case of Macro-Mayan and most importantly, in the case of Nostratic family, the sub-grouping of Indo-European and Dravidian languages.

Acknowledgements

To work on a dissertation, whose long term goal as atypical as attempting to unearth a relationship between Indo-European and Dravidian, unsurprisingly could not have been a smooth journey. In this respect, the late Prof. Stephan H. Levitt (1943-2020) once wrote about the discouragement he had encountered when he had placed this topic in front of a few scholars (Levitt, 1998). I, on the other hand, was given room by my advisor to express ideas and was given sufficient encouragement by him to pursue this unconventional, yet highly relevant topic as part of my dissertation. For this, I would like to thank my supervisor Prof. Arnab Bhattacharya for his guidance and support throughout my PhD. I would also like to thank him along with Pralay Manna Mahodaya for organizing Sanskrit classes at IIT Kanpur, which served as an entry point for me to linguistic studies. I am also grateful to my advisor for organizing sessions aimed towards multilingualism which deepened my knowledge in various Indic languages, ultimately providing me impetus to pursue comparative historical linguistics. In this regard, I extend my gratitude to my friends and colleagues, Dr. Hrishikesh R. Terdalkar, Dr. Shubhangi Agarwal, Dr. Chaitali Dangarikar, Pramit Bhattacharyya, Shubham K. Nigam and Dr. Oswald Christopher, who were part of those sessions and from whom I gained insights into Indic languages through memorable discussions. I would also like to thank Dr. Chinmay Dharurkar for involving me in discussions that improved my linguistic understanding. I further owe my thanks to Priya and Debkanta Chakraborty for helping me understand computational phylogenetics through fruitful discussions. My deep gratitude to all my teachers at IIT Kanpur, especially to those who encouraged me to pursue a research career. I would also like to thank the Ministry of Education of India (MHRD) for funding this research through Prime Minister's

Research Fellowship (PMRF).

My primary source of inspiration and confidence to study Sanskrit, Old Telugu and, thus, linguistics has been from the profound articles and books of Sri Pandita Bandaru Tammayya (1891-1970). I would like to thank his granddaughter, my mother for preserving his books and passing them over to me. I would also like to acknowledge my family members, including my grandmother, father, sister, late grandfathers, and others, who have been influential in my acquisition of traditional Telugu. Finally, I express my deep gratitude to my spiritual preceptor, who directed me to focus particularly on Śaivite texts in Old Telugu. My journey through linguistics so far would not have been possible without these foundations, for which I am grateful to all those involved.

To the Śaivite and the Vaiṣṇavite saints,

who sung in the Old Dravidian languages

Contents

List of Tables

List of Figures

Chapter 1

Introduction

āriyamun-tamiḻum uṭanē collik-kārikaiyārkkuk-karuṇai ceytāṉē

"Then did He in Sanskrit and Tamil at once, Reveal the rich treasure of His Compassion to our Lady Great.[1] "

– *Tirumantiram* 65, Tirumūlar, ∼400-800 CE, Tiruvāvaṭuturai

1.1 Problem Statement

Languages are prone to change both phonologically and grammatically with time, leading to diversification across geographical spans, thereby subsequently giving rise to dialects or ultimately new languages. In this process, languages form groups of "language families" or *phylogenies* which reflect their ancestry. Languages that arise from a common ancestral language are termed to be *genetically related*, and they often carry the signatures of common ancestry in the form of lexical resemblances or *cognates*. *Historical linguistics* is a discipline that aims to study such diachronic evolution of languages. Typical tasks include identifying cognates, reconstructing ancestral languages or *proto-languages*, classification of languages into language families, etc. (Campbell, 2013). These tasks often demand a great amount of manual effort. Thus, there is a need for automation to assist the historical linguists.

[1] English Translation from *Tirumantiram*, Sri Ramakrishna Math, Mylapore, Chennai

The field of *computational historical linguistics*, thus, aims to alleviate the human labor that goes into historical linguistics. Owing to the similarities with evolutionary biology, this field significantly borrows several concepts and techniques such as sequence alignment, phylogenetic inference, and others from computational biology and bioinformatics (Jäger, 2019). These techniques in combination with those of natural language processing, string comparison, and machine learning are employed for the problems of historical linguistics such as significance testing of genetic relationship (Oswalt, 1970; Ringe, 1992, 1996; Kessler, 2001; Kessler and Lehtonen, 2006; Kessler, 2007), phonemic sequence alignment (Kondrak, 2000; Bhargava and Kondrak, 2009; List, 2010), automated cognate detection (List, 2010, 2012b,a; Rama, 2016; Jäger and Sofroniev, 2016; Jäger et al., 2017; List et al., 2017; Rama and List, 2019; MacSween and Caines, 2020), proto-language reconstruction (Bouchard-Côté et al., 2013; Ciobanu and Dinu, 2018; Meloni et al., 2021; List et al., 2022a,b; Kim et al., 2023), phylogenetic inference (Jäger, 2015, 2018; Rama et al., 2018; Rama and List, 2019) and several others.

Among the problems stated above, this thesis addresses the problems of automated cognate detection, automated phonological reconstruction, and significance testing of genetic relationships. In **cognate detection**, cognates, i.e., words that descend from a common ancestral word are identified from multilingual wordlists across each concept or meaning. On the other hand, in **phonological reconstruction** or simply, proto-language reconstruction, the proto-word forms of an ancestral language are reconstructed based on the attested cognates among the daughter languages. As an illustration, consider some words in a few Indo-European languages for the concept 'I', namely, Sanskrit *ahám*, Greek *egó*, Latin *egō*, German *ich*, and Lithuanian *aš*, which are cognates. The corresponding proto-form[2] for this meaning in ancestral Proto-Indo-European can be reconstructed as **egʰ₂óm*. A significance test, as the name suggests, is employed to check whether lexical resemblances among a group of languages are chance occurrences or not, henceforth suggesting whether the given group of languages are genetically related or not.

The existing methods addressing these problems come with certain limitations listed as

[2]For understanding the notation of Proto-Indo-European please see the page explaining its phonology: https://en.wikipedia.org/wiki/Proto-Indo-European_phonology.

follows for each problem. These are addressed by the methods we propose in this thesis.

- In *proto-language reconstruction*, transfer learning, i.e., transfer of knowledge from known data to unknown, has been unexplored since the existing methods tend to train separately on a per-language family basis.

- In *cognate detection*, existing methods are mostly unsupervised. Thus, the labeled cognacy information is under-utilized, which could potentially be useful.

- In *significance testing*, we observed that the existing multilateral permutation tests may yield a false positive on an unrelated language group. Other existing tests rely on proto-language reconstructions. However, historical linguists often disagree on the reconstructed proto-forms.

This thesis proposes novel methods to address these challenges employing the representational power of a transformer (Vaswani et al., 2017) in combination with ideas from the developments in computational biology such as in protein fold prediction (Rao et al., 2021; Jumper et al., 2021) or in significance testing of phylogenetic aspects (Huelsenbeck and Bull, 1996; Huelsenbeck et al., 1996; Anisimova and Gascuel, 2006). The methods thus proposed achieve state-of-the-art results on the above-stated problems. The aims of this thesis are mentioned in the following section.

1.2 Aims and Objectives

The aims of the research conducted as part of this thesis are listed as follows:

- To develop computational tools for the three aforementioned problems, namely automated phonological reconstruction, automated cognate detection, and significance testing of genetic relatedness, that can overcome the limitations of existing methods.

- To maintain an integration of biological insights throughout following the tradition of computational historical linguistics.

- To evaluate the performances of the proposed methods against those of existing methods especially at low resource settings.
- To maintain multilingualism, i.e., to ensure language diversity through out the experiments.
- To bear broader implications, especially by applying the significance tests on proposed macro-families.

The contributions of this thesis fulfilling the above aims are summarized in the following section.

1.3 Contributions

The contributions of this thesis are listed as follows:

- We introduce **Cognate Transformer (CogTran)** based on a protein language model (Rao et al., 2021) that performs comparable to or better than the existing methods on the tasks of proto-language reconstruction and cognate reflex prediction.
- We also demonstrate the advantage of pre-training CogTran on the task of masked reflex prediction (akin to masked language modeling).
- We propose a supervised method named **CogTran2**, for automated cognate detection that outperforms previous methods in the presence of little additional supervision. Thus, it utilizes labeled data unlike the previous methods. The model can adapt and perform well on new language families with fine-tuning on only a few concepts (akin to few-shot learning).
- CogTran2 is also efficient in terms of time in comparison with previous methods, since it consists of an end-to-end architecture that directly takes multiple sequences as inputs to output cognate cluster linkages avoiding pairwise instantiations.
- We incorporate modules as part of CogTran2 to capture transitive property on output cognate clusters which is responsible for better performance.

- We propose a **likelihood ratio test (LRT)** to determine the genetic relatedness of a group of languages based on the proportions of phonetically invariant sites.
- We find that LRT does not exhibit the problems of false positives nor does it require proto-forms, which are the issues with previous methods.
- Finally, by application of LRT and other tests, we find supporting evidence for sub-groupings Indo-European-Dravidian and Mayan-Mixe-Zoque of macro-families Nostratic and Macro-Mayan respectively.

1.4 Organization of Thesis

The rest of the thesis is organized as follows. Background knowledge on computational historical linguistics is given in Chapter 2. The aims and objectives of our research are mentioned in Chapter 3. Cognate Transformer is introduced in Chapter 4in the context of proto-language reconstruction. Its adaptation to cognate detection by means of additional modules supporting linkage prediction is described in Chapter 5. A significance test of genetic relationship, namely the likelihood ratio test is described in Chapter 6, along with its application to the testing of a few macro-families. The thesis is finally concluded in Chapter 7 along with the discussions on potential future work.

1.5 Published Material

The core chapters of this thesis have been either published or yet to appear in conference proceedings. The details are mentioned as follows:

- Chapter 4 is based on the work that was published in the proceedings of 2023 conference of the Empirical Methods in Natural Language Processing (Akavarapu and Bhattacharya, 2023a).
- Chapter 5 is based on the work that was published in the proceedings of the 18th conference of the European Chapter of the Association of Computational Linguistics (Akavarapu and Bhattacharya, 2024a).

- Chapter 6 is based on the work that was published in the proceedings of the 2024 conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies (Akavarapu and Bhattacharya, 2024b).

The work Akavarapu and Bhattacharya (2023b), although very different in application, has been inspirational from a methodological point of view.

Chapter 2

Background on Computational Historical Linguistics

This chapter gives an introduction to the field of computational historical linguistics and mentions the major advances in this field. Before proceeding to it, the following section gives a brief introduction to historical linguistics.

2.1 Historical Linguistics

Historical linguistics is the study of language change across time (Campbell, 2013). In other words, it is the study of how a language evolves from its ancestral language, how it differs in the course of evolution from its sister languages, how it gets shaped by the influence of languages that come into its contact, how it gets diversified into various dialects or even newer languages, and in many cases, how it may face extinction. Such language change is studied in terms of phonetic, morphological, and syntactical changes.

The earliest systematic study of language change can be attributed to the Prakrit grammarians (Subrahmanyam, 2011), a fact neglected by historical linguists. For instance in the Prākṛta Prakāśa of Vararuci, one can find several rules describing phonetic and morphological changes from Sanskrit to various Prakrit dialects. The beginnings of modern comparative historical linguistics, however, are attributed to a passage by Sir Willian Jones, a philologist, delivered before the Asiatic Society at Calcutta in 1786, given

as follows:

"The Sanscrit language, whatever be its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either, yet bearing to both of them a stronger affinity, both in the roots of verbs and the forms of grammar, than could possibly have been produced by accident; so strong indeed, that no philologer could examine them all three, without believing them to have sprung from some common source, which, perhaps, no longer exists; there is a similar reason, though not quite so forcible, for supposing that both the Gothic and the Celtic, though blended with a very different idiom, had the same origin with the Sanscrit; and the old Persian might be added to the same family." (Jones, 1824)

In this passage, Jones narrates his findings where he observed Sanskrit to be similar to Greek and Latin and concluded that they should have originated from the same source language which is now referred to as Proto-Indo-European. This passage also gives a glimpse of what is known as the comparative method, a central methodology of historical linguistics that involves systematic comparisons of languages to inquire about a possible genealogical relationship among them. This method is discussed in more detail as follows.

2.1.1 The Comparative Method

The comparative method aims to reconstruct the linguistic past through systematic comparisons of related languages. The steps involved are gathering cognates, identifying regular sound correspondences, reconstruction of the proto-forms, reconstruction of morphemes, and classification of the languages into language families/sub-families (Campbell, 2013). Languages in a language family are termed to be *genetically related* due to their descendence from a common ancestral language. As an illustration, consider the preposition/prefix cognate words meaning 'under' namely Sanskrit *upá*, Ancient Greek *hypó* and Latin *sub* where the meaning 'under' can be inferred from modern words such as (Sanskrit) ***upā****dhyakṣa* 'vice-president' or 'deputy-chairman', ***hypo****glycemia* 'having low blood sugar' or ***sub****marine* literally 'under-sea'. Identification of sound correspondences, in this example, refers to matching the consonants *p-p-b* or the root vowels *u-y-u* respectively

in Sanskrit, Greek, and Latin. The corresponding ancestral proto-form is reconstructed as **upó* in Proto-Indo-European taking into account observed sound correspondences from other cognate sets. An example of a language sub-group would be that of Romance languages, the languages that derive from Vulgar Latin. Another example would be that of Indo-Aryan consisting of languages that descend from (Vedic) Sanskrit and are concentrated in the Northern part of India. Finally, the evolution history of a language family is represented by an evolutionary tree also referred to as a *phylogeny*.

The tasks of the comparative method are traditionally carried out manually involving strenuous efforts. While it is still feasible for small to medium-sized language families, the process becomes almost infeasible for large language families. For example, consider the Austronesian family that consists of over a thousand languages geographically spanning a vast area from Madagascar in Africa to Easter Island near Chile or up to the Hawaiian islands. Further, each language in the family would have thousands of words. Hence, it is desirable to have automated tools to assist historical linguistics in carrying out the comparative method on such large language families.

Before talking about automating historical linguistics, it is worth noting the similarities between language evolution and biological evolution described as follows.

2.1.2 Parallels with Evolutionary Biology

The existence of analogies between the evolutionary processes of languages and species was noted by Darwin himself in *The Descent of Man* (1871). At a fundamental level, the phonemic sequence across the lexica of a language is analogous to the DNA sequence of an organism which is inherited by its descendants and which may change, i.e. mutate at certain sites leading to diversification. The evolutionary history in both cases can be, thus, represented in the form of a tree known as a *phylogeny*. These fundamental analogies consequently give rise to several others, some of which are listed in Table 2.1 (Atkinson and Gray, 2005; Campbell, 2013). For instance, consider horizontal gene transfer which refers to the movement of genetic information across distantly related species (Keeling and Palmer, 2008). In linguistics, on the other hand, one finds the analogous phenomenon

Biological evolution	Linguistic evolution
Discrete characters	Lexicon, syntax, and phonology
Homologies	Cognates
Mutation	Sound change / Innovation
Natural selection	Social selection
Cladogenesis	Diversification
Horizontal gene transfer	Borrowing / Language Contact
Plant hybrids	Language Creoles
Correlated genotypes-phenotypes	Correlated cultural terms
Geographic clines	Dialects / Dialect chains
Fossils	Relics / Archaisms
Extinction	Language death

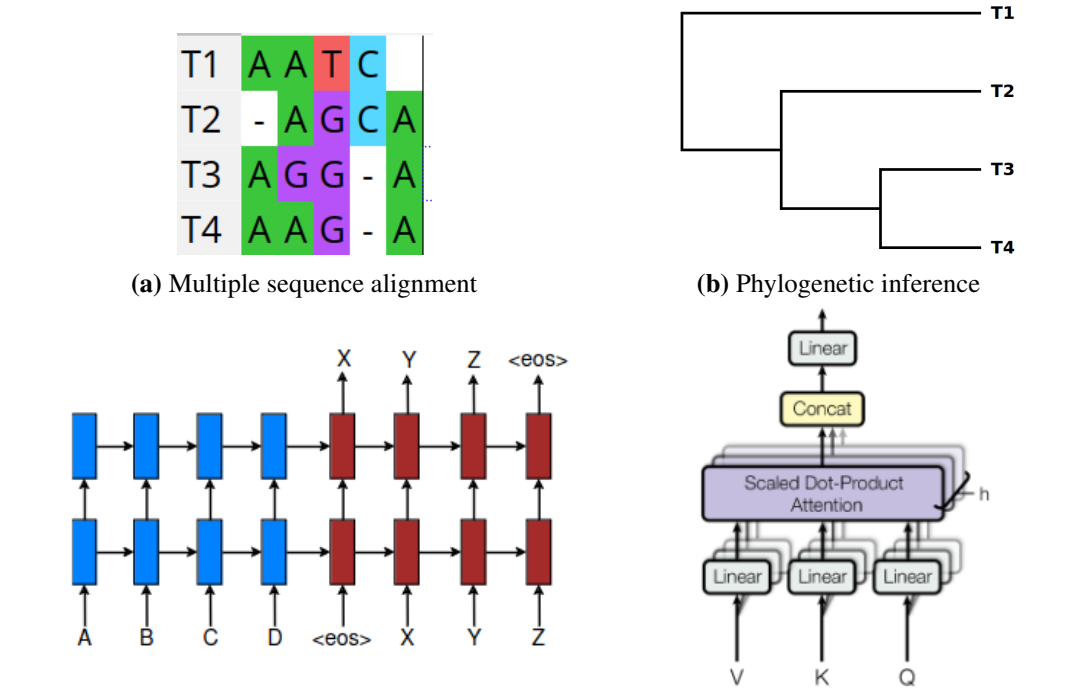
Table 2.1: Parallels between the evolutionary processes of living organisms and languages

of borrowing. For example, English word *orange* arrived by passing through various languages ultimately from a Dravidian source related to Tamil *nāram* and Telugu *nāriñja*.

The presence of such analogies made it possible to apply directly many methods developed for evolutionary biology to linguistic data such as those of computational phylogenetics. This led to the significant emergence of computational historical linguistics, which is discussed in the following section.

2.2 Computational Historical Linguistics

As discussed in §2.1.1, there is a need to automate the tasks of the comparative method to assist the historical linguistics to facilitate easy handling of large language families with thousands of languages. Computational historical linguistics, a relatively newer field, aims to this end. Although this field existed in the previous century, it received a significant boost since the early 2000s due to primarily importing ideas and techniques from computational biology (Jäger, 2019). This progress owes to the presence of various analogies between historical linguistics and evolutionary biology that have been just briefly touched upon (§2.1.2). Computational historical linguistics shares insights and techniques with three traditional disciplines, namely the comparative method, computational biology / bioinformatics, and computational linguistics / natural language processing (Jäger, 2019). These are elaborated further for each discipline as follows.



(a) Multiple sequence alignment

(b) Phylogenetic inference

(c) Neural machine translation (Luong et al., 2015)

(d) Multihead attention (Vaswani et al., 2017)

Figure 2.1: Computational historical linguistics borrows heavily from computational biology (top row) and computational linguistics (bottom row). Multiple sequence alignment and phylogenetic inference are performed in MEGA11 (Tamura et al., 2021)

2.2.1 Relation to the Comparative Method

Automating the tasks of the comparative method is the primary goal of this field. Some of the central topics include automatic assessment of genetic relatedness, automatic cognate detection, phylogenetic inference, and ancestral state reconstruction, i.e., proto-language reconstruction (Jäger, 2019). These are verily the problems that are addressed in this thesis respectively in chapters 6, 5 and 4. Refer to the examples from §2.1.1 for the notions of cognates, genetic relatedness, phylogeny, and ancestral state reconstruction.

2.2.2 Relation to Computational Biology

Early algorithms inspired by computational biology in this field had to do with phonetic alignment, for instance, dynamic programming based ALINE (Kondrak, 1999, 2000). To this day, widely used phonetic alignment modules from LingPy library (List and Forkel, 2021) make use of the pairwise molecular sequence alignment algorithms by

Needleman and Wunsch (1970), and Smith and Waterman (1981). Sequence alignment is fundamental to most problems such as cognate detection, proto-language reconstruction, or phylogenetic inference. All the methods in this thesis indeed begin from multiple sequence alignments. Another work that notably builds from evolutionary biology is that of Bouchard-Côté et al. (2013), where a probabilistic model was proposed for proto-language construction akin to ancestral state reconstruction of nucleotide or amino acid sequences. Prominent import from computational biology apart from sequence alignment has been that of phylogenetic inference (Jäger, 2019). It is not uncommon to date to employ some phylogenetic software developed for molecular or trait data on cognacy data. For instance in this thesis (Chapter 6), IQ-TREE (Nguyen et al., 2015), a phylogenetic software, was used to compute maximum-likelihood trees as part of phylogenetic testing. Another commonly used package is MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist et al., 2012) for Bayesian phylogenetic inference as applied in, for instance, Rama et al. (2018) or Jäger (2019).

Other class of methods that are inspired by computational biology concern with signficance testing of genetic relatedness. Many are based on Fischer's permutation test such as that of Oswalt (1970) or Kessler (2001); Kessler and Lehtonen (2006) which are often found in applications to biological sequences (Doerge and Churchill, 1996; Faith, 1991). In this thesis (Chapter 6), likelihood ratio test was used for this problem. Likelihood ratio test is often applied to test hypotheses on phylogenies as in case of Huelsenbeck and Bull (1996); Huelsenbeck et al. (1996) and Anisimova and Gascuel (2006). Other tests of genetic relatedness use brute-force probability computations on sound correspondences and are not biologically inspired (Ringe, 1992, 1996). However, these methods have certain theoretical problems with the probability computations as pointed out by Kessler (2001) and are thus, considered obsolete.

2.2.3 Relation to Computational Linguistics

This field of computational historical linguistics is ultimately part of the broader computational linguistics and thus, advances therein are passed on or inherited here. For a

basic example, string comparison metrics like edit distances are used as a subroutine to compute distances among lexicon in cognate detection task (List et al., 2017). Since biological sequences are also ultimately strings, there is a significant overlap of methodology in this respect as well. For instance, specific types of hidden Markov models found in bioinformatics have been applied to the problems of cognate detection (Mackay and Kondrak, 2005) and to the multiple sequence alignment problem (Bhargava and Kondrak, 2009). Application of statistical methods to proto-language reconstruction apart from that of Bouchard-Côté et al. (2013) includes conditional random fields based on Latin word reconstruction by Ciobanu and Dinu (2018). Recent methods include neural machine translation-based proto-language reconstruction (Meloni et al., 2021) and ultimately with the transformers (Celano, 2022; Kim et al., 2023). Some methodologies of computational historical linguistics shared with computational biology and computational linguistics are illustrated in Figure 2.1. In this thesis, we have employed transformer-based architectures based on MSA Transformer (Rao et al., 2021) and AlphaFold2 (Jumper et al., 2021) that were originally applied to amino acid sequences to the problems of proto-language reconstruction (Chapter 4) and to that of cognate detection (Chapter 5). Here, it should be emphasized although these models come from computational biology, the underlying transformer architecture (Vaswani et al., 2017) or the underlying attention principle (Bahdanau et al., 2014) ultimately arose in computational linguistics in the context of neural machine translation.

This completes the brief introduction to both computational historical linguistics and its current state. The existing approaches in this field, however, do come with limitations and challenges, some of which are mentioned as follows.

2.2.4 Limitations of the Existing Methods

Here, we discuss the limitations of the existing methods concerning the problems addressed in this thesis viz., proto-language reconstruction, cognate detection, and significance testing of genetic relationship. These are elaborations of the ones listed in Chapter 1 (§1.1).

- In the problem of proto-language reconstruction, the existing methods are trained

separately for different language families. Although every language family may have a distinct set of sound changes defining its evolution, the possible sound changes will still have great overlap across families. The reason behind this has to do with the fact that most sound changes are universal partly due to the anatomy of speech production. Thus, one may consider 'transfering' the common knowledge i.e., of the sound changes learned from a group language families to others, which is not observed with the existing methods.

- In the problem of cognate detection, the existing methods are mostly based on unsupervised computation of certain word similarity measures that are further used to cluster out cognates. However, there is a great amount of information on cognacy labels across several language families for at least about a hundred meanings which is not utilized. Thus, again one may consider a supervised model that can be trained on such data and ideally, which should be adaptable to unseen families without further significant efforts.

- In the problem of significance testing of genetic relatedness, existing methods that are based on permutations of multilateral wordlists (Kessler, 2007), as shall be shown in Chapter 6, exhibit false positives, i.e., may sometimes consider an unrelated group of languages as related. Other methods, on the other hand, rely on the reconstructed proto-forms (Ringe, 1996; Kassian et al., 2015) whose reconstructions are often a point of contention among the historical linguists.

The methods proposed in this thesis aim to address these limitations and further advance the state-of-the-art.

2.3 Summary

Thus, this chapter briefly introduces the subfield of computational historical linguistics which lies at the intersection of three fields namely historical linguistics, computational biology, and computational linguistics. The problems addressed in this thesis namely, proto-language reconstruction, cognate detection, and significance testing of genetic relatedness

have also been discussed along with the existing methods and their limitations. In this thesis, we propose novel methods for these problems following the general theme of computational historical linguistics, i.e., being situated at the triple confluence of historical linguistics, computational biology, and computational linguistics, as it becomes evident through the following chapters.

Chapter 3

Aims and Objectives

In this chapter, the aims and objectives of the research conducted as part of this thesis are explained which have been briefly mentioned in Chapter 1 (§1.2). The common goals and methodologies that knit together the subsequent chapters to advance the state-of-the-art align well with those of computational historical linguistics as seen in the previous Chapter 2 namely, to automate the tasks of the comparative method employing insights and methods from computational biology and computational linguistics. These are reiterated in this chapter as it shall become evident through the aims and objectives listed as follows.

3.1 Development of Computational Tools

The primary aim of this thesis is to advance the state-of-the-art computational tools that perform proto-language reconstruction, cognate detection, and significance testing of genetic relatedness. To this end, the following objectives were set addressing the limitations of the existing methods (see §2.2.4):

1. To develop a neural architecture, in our case, a transformer-based model for proto-language reconstruction that can be trained on datasets from multiple language families, thereby capable of transfer learning by design. In other words, a model, which is once sufficiently 'pre-trained', can be fine-tuned easily on an unseen low-resource language family's data. To this end, we developed the Cognate Transformer (CogTran) described in Chapter 4.

2. To develop a supervised approach by adapting the above model to the problem of cognate detection so that the labeled information can be utilized efficiently. Ideally, the trained model should be adaptable to unseen families without further significant fine-tuning. Since, the problem of cognate detection involves clustering of word pairs (see §5.3), a supervised algorithm can be designed by viewing the problem as a link prediction problem where links or edges should be defined to exist among the members of the same cluster. To this end, we developed CogTran2 described in Chapter 5.

3. To develop a significance test of genetic relatedness that can simultaneously be able to test on multiple languages akin to multilateral permutation tests (§6.4.2), at the same time avoid the false positives exhibited by the latter (§6.5). To this end, we proposed a likelihood ratio test (LRT) described in Chapter 6.

3.2 Integration of Biological Insights

The thesis also aims to retain the tradition of adapting methods from computational biology in computational historical linguistics, an aspect that has been emphasized in the previous chapter. Objectives to achieve this are to make use of an aligned sequence structure of cognates and to employ models that can utilize this inherent structure. To this end, all three methods proposed in this thesis begin from multiple sequence alignment (MSA) input, followed by sub-routines apt to such input. For instance, CogTran and CogTran2 adapt from a protein language model (Rao et al., 2021) and a protein structure predictor (Jumper et al., 2021), while LRT depends on phylogenetic tree construction and is inspired by hypothesis testing methods on phylogenetic trees of molecular phylogenetics.

3.3 Performance Evaluation at Low Resource Settings

There are several thousands of languages whose cognacy data is either unavailable or very scarce. Hence, it is a basic requirement of the methods in this field to be able to work well

in such low-resource settings. Thus, the thesis aims to evaluate the performance of the proposed methods both qualitatively and quantitatively against that of the existing methods, especially in low-resource settings. To this end, CogTran and CogTran2 both outperform the existing methods across all test proportions, especially at low resource settings.

3.4 Mutilinguality

The common aim of any natural language processing algorithm should be to be able to work well ideally on any language to serve all the languages of the world alike. Hence, this thesis aims to apply the developed methods to the datasets from as many phonetically diverse language families as possible. To this end, the datasets used in all the problems reflect language diversity.

3.5 Broader Implications

Finally, this thesis aims to bear implications, some of which, perhaps important. Firstly, in terms of providing better assistance to the historical linguists through refined methods. Secondly and importantly, applications of significance tests (of Chapter 6) on potential newer language families namely Macro-Mayan and Nostratic. Such statistical evidence would provide impetus to look for finer linguistic similarities shared by the languages in each group which in turn provide crucial insights into the past of the civilizations that spoke these languages. To this end, significance tests applied on macro-families support genetic relatedness of Mayan-Mixe-Zoquen and Indo-European-Dravidian.

The subsequent chapters present the methodologies that aim to fulfill the aims and objectives set in this chapter.

Chapter 4

Cognate Transformer for Phonological Reconstruction

Phonological reconstruction is one of the central problems in historical linguistics, as described in Chapter 2, where a proto-word of an ancestral language is determined from the observed cognate words of daughter languages. Computational approaches to historical linguistics attempt to automate the task by learning models on available linguistic data. Several ideas and techniques drawn from computational biology have been successfully applied in the area of *computational historical linguistics*, as emphasized in Chapter 2. Following these lines, we adapt MSA Transformer, a protein language model, to the problem of *automated phonological reconstruction* in this chapter. MSA Transformer trains on multiple sequence alignments as input and is, thus, apt for application on aligned cognate words. We, hence, name our model as *Cognate Transformer*. We also apply the model on another associated task, namely, *cognate reflex prediction*, where a reflex word in a daughter language is predicted based on cognate words from other daughter languages. Finally, we show that our model outperforms the existing models on both tasks, especially when it is pre-trained on masked word prediction task. This chapter is based on Akavarapu and Bhattacharya (2023a).

4.1 Introduction

Phonological reconstruction of a word in an ancestral proto-language from the observed cognate words, i.e., words of supposed common origin, in the descendant languages is one of the central problems in *historical linguistics*, a discipline that studies diachronic evolution of languages (Campbell, 2013). For example, the cognate words French *enfant*, Spanish *infantes* and Italian *infanti* all trace to the proto-form *infantes* in Latin meaning 'children', which is an attested language in this case. In most cases, the proto-language is not attested and has to be rather reconstructed. The process of arriving at such phonological reconstruction usually involves multiple steps including gathering potential cognate words, identifying systematic sound correspondences, and finally reconstructing the proto-phonemes. This procedure is known as the 'comparative method' (Ringe and Eska, 2013), which is traditionally carried out manually.

Several *automated phonological reconstruction* algorithms emerged in the last decade. Some of these are drawn or inspired from computational biology, for example, Bouchard-Côté et al. (2013). In general, computational historical linguistics draws techniques such as sequence alignment and phylogenetic inference from computational biology, in addition to the techniques known from historical linguistics and computational linguistics or natural language processing (Jäger, 2019). On similar lines, we adapt the MSA transformer, introduced in Rao et al. (2021) for modeling multiple sequence alignment (MSA) protein sequences, for the problem of phonological reconstruction which takes as input a cognate word set in the form of MSAs. Henceforth, we name the model introduced here as *Cognate Transformer*.

We also apply our model on *cognate reflex predicion* task, where an unknown, i.e., a masked reflex in a daughter language is to be predicted based on the attested reflexes from other daughter languages (List et al., 2022b). For instance, in the previous example, if we mask French *enfant*, the task would involve arriving at the word form correctly based on Spanish *infantes* and Italian *infanti*. One can notice that this task can serve as a pre-training objective for the proto-language reconstruction task described previously. Hence, we also

pre-train the Cognate Transformer on the cognate reflex prediction task.

Further, most of the existing models are fitted on a per language family basis, i.e., on one dataset at a time consisting of a single language family. Thus, the utility of either transfer learning or simultaneous fitting across several language families has not yet been demonstrated. This is desirable even from the linguistic perspective since it is well known that the sound changes are phonologically systematic and, thus, often similar sound changes operate across different language families (Campbell, 2013). For instance, the sound change involving palatalization of a velar consonant say /k/ > /tʃ/ can be observed in the case of Latin *caelum* /kaɪlʊm/ to Italian *cielo* /tʃɛːlo/ as well as in the supposed cognate pairs *cold* versus *chill*, which is a reminiscence of historical palatalization in Old English.[1] Hence, owing to the presence of commonalities across language families in terms of sound change phenomena, training models simultaneously across multiple language families should be expected to yield better results than when training on a single language family data at a time. This is well reflected in our present work.

4.1.1 Problem Statements

There are two tasks at hand as mentioned before, namely, *cognate reflex prediction* and *proto-language reconstruction*.

An input instance of the cognate reflex prediction task consists of a bunch of cognate words from one or more related languages with one language marked as unknown; the expected output would be the cognate reflex in that particular language which is marked unknown. An example from the Romance languages is:

Input: [French] ʒ ə n j ɛ v ʁ,
[Portuguese] ?,
[Italian] dʒ i n e p r o

Output: [Portuguese] ʒ u n i p ɨ ɾ ʊ

The input for the proto-language reconstruction task consists of cognate words in the daughter languages and the expected output is the corresponding word in the ancestral

[1]For International Phonetic Alphabet (IPA) notation, see https://en.wikipedia.org/wiki/International_Phonetic_Alphabet.

(proto-) language. We model this as a special case of cognate reflex prediction problem where the proto-language is always marked as unknown. For instance, in the following example, Latin would be marked as unknown:

Input: [Latin] ?,
[French] ʒ ə n j ɛ v ʁ,
[Portuguese] ʒ u n i p ɨ ɾ ʊ
Output: [Latin] j uː n ɪ p ɛ r ʊ m

4.1.2 Contributions

Our contributions are summarized as follows. We have designed a new architecture, Cognate Transformer, and have demonstrated its efficiency when applied to two problems, namely, proto-language reconstruction and cognate reflex prediction, where it performs comparable to the existing methods. We have further demonstrated the use of pretraining in proto-language reconstruction, where the pre-trained Cognate Transformer outperforms all the existing methods.

The rest of the chapter is organized as follows. Existing methodologies are outlined in §4.2. The workflow of Cognate Transformer is elaborated in §4.3. Details of experimentation including dataset information, model hyperparameters, evaluation metrics, etc. are mentioned in §4.4. Results along with discussions and error analysis are stated in §4.5.

4.2 Related Work

Several methods to date exist for proto-language reconstruction, as mentioned previously. We mention a notable few. Bouchard-Côté et al. (2013) employs a probabilistic model of sound change given the language family's phylogeny, which is even able to perform unsupervised reconstruction on Austronesian dataset. Ciobanu and Dinu (2018) performed proto-word reconstruction on Romance dataset using conditional random fields (CRF) followed by an ensemble of classifiers. Meloni et al. (2021) employ GRU-attention based neural machine translation model (NMT) on Romance dataset. List et al. (2022a) presents datasets of several families and employs SVM on trimmed alignments.

The problem of cognate reflex prediction was part of SIGTYP 2022 shared task (List et al., 2022b), where the winning team (Kirov et al., 2022) models it as an image inpainting problem and employs a convolutional neural network (CNN). Other high performing models include a transformer model by the same team, a support vector machine (SVM) based baseline, and a Bayesian phylogenetic inference based model by Jäger (2022). Other previous approaches include sequence-to-sequence LSTM with attention, i.e., standard NMT based (Lewis et al., 2020) and a mixture of NMT experts based approach (Nishimura et al., 2020).

The architecture of MSA transformer is part of Evoformer used in AlphaFold2 (Jumper et al., 2021), a protein structure predictor. Pre-training of MSA transformer was demonstrated by Rao et al. (2021). Handling MSAs as input by using 2D convolutions or GRUs was demonstrated by Mirabello and Wallner (2019) and Kandathil et al. (2022).

4.3 Methodology

In this section, the overall workflow is described. The input phoneme sequences are first aligned (§4.3.1), the resulting alignments are trimmed (§4.3.2), and then finally passed into the MSA transformer with token classification head (§4.3.3). In the training phase, the output sequence is also aligned while in the testing phase, trimming is not performed. The first two steps are the same as described in List et al. (2022a) and are briefly described next.

4.3.1 Multiple Sequence Alignment

The phenomenon of sound change in spoken languages and genetic mutations are similar. As a result, multiple sequence alignment and the methods surrounding it are naturally relevant here as much as they are in biology. The phonemes of each language in a single cognate set are aligned based on the sound classes to which they belong. An example of an alignment is given in Table 4.1.

We use the implementation imported from the library `lingpy` (List and Forkel, 2021)

[French]	ʒ	ə	n	j	ɛ	v	-	ʁ	-	-
[Italian]	dʒ	i	n	-	e	p	-	r	o	-
[Spanish]	x	u	n	-	i	p	e	ɾ	o	-
[Latin]	j	uː	n	-	ɪ	p	ɛ	r	ʊ	m

Table 4.1: Aligned phoneme sequences

[French]	ʒ	ə	n	j.ɛ	v	-	ʁ	-
[Italian]	dʒ	i	n	e	p	-	r	o
[Spanish]	x	u	n	i	p	e	ɾ	o
[Latin]	?	?	?	?	?	?	?	?
[Latin]	j	uː	n	ɪ	p	ɛ	r	ʊ.m

Table 4.2: Trimmed input and output alignments

which uses the sound-class-based phonetic alignment described in List (2012b). In this algorithm, the weights in pairwise alignments following Needleman and Wunsch (1970) are defined based on the sound classes into which the phonemes fall. Multiple sequences are aligned progressively following the tree determined by UPGMA (Sokal and Michener, 1975).

4.3.2 Trimming Alignments

In the example given in Table 4.1, one can observe that during testing, the final gap (hyphen) in the input languages (i.e., excluding Latin) will not be present. Since the task is essentially a token classification, the model will not predict the final token 'm' of Latin. To avoid this, alignments are trimmed as illustrated in Table 4.2 for the same example.

This problem is discussed in detail in (List et al., 2022a) and the solution presented there has been adopted here. In particular, given the sequences to be trimmed, if in a site all tokens are gaps except in one language, then that phoneme is prepended to the following phoneme with a separator and that specific site is removed. For the last site, the lone phoneme is appended to the penultimate site. Following (List et al., 2022a), trimming is skipped for testing as it has been observed to cause a decrease in performance. The reason for this is mentioned in (Blum and List, 2023). Briefly stating it, gaps in daughter languages can point to a potential phoneme in the proto-language. While training however,

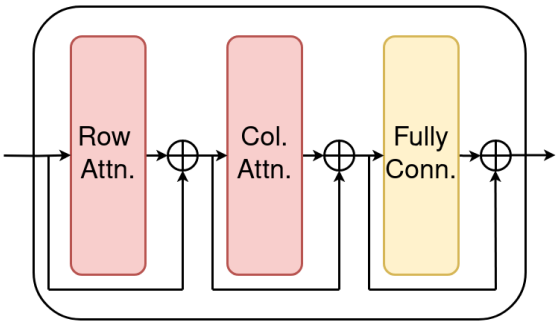


Figure 4.1: A single layer of MSA Transformer.

they are redundant and can be trimmed since proto-language is part of alignment.

4.3.3 MSA Transformer

The MSA Transformer, proposed in (Rao et al., 2021), handles two-dimensional inputs with separate row and column attentions (each with multiple heads) in contrast with the usual attention heads found in standard transformer architectures (Vaswani et al., 2017). It uses learned positional embeddings only across rows since a group of rows does not make up any sequential data. The outputs of row attentions and column attentions are summed up before passing into a fully connected linear layer (see Figure 4.1). MSA Transformer, despite its name, is not an encoder-decoder transformer but rather only an encoder like BERT (Devlin et al., 2018), except with the ability to handle 2D input (see Figure 4.2). For more information on row and column attentions see Appendix A.

4.3.4 Workflow

The aligned input sequences thus trimmed are passed into MSA Transformer as tokens. A single input instance to an MSA Transformer is a 2D array of tokens. The overall architecture of the Cognate Transformer is illustrated in Figure 4.2. Due to trimming, several phonemes can be joined together as one token. Hence, with trimming the total number of tokens or the vocabulary size can be above 1000 or even 2000 based on the training dataset, while without trimming the vocabulary size would essentially be close to the total number of phonemes possible which would be only a few hundreds.

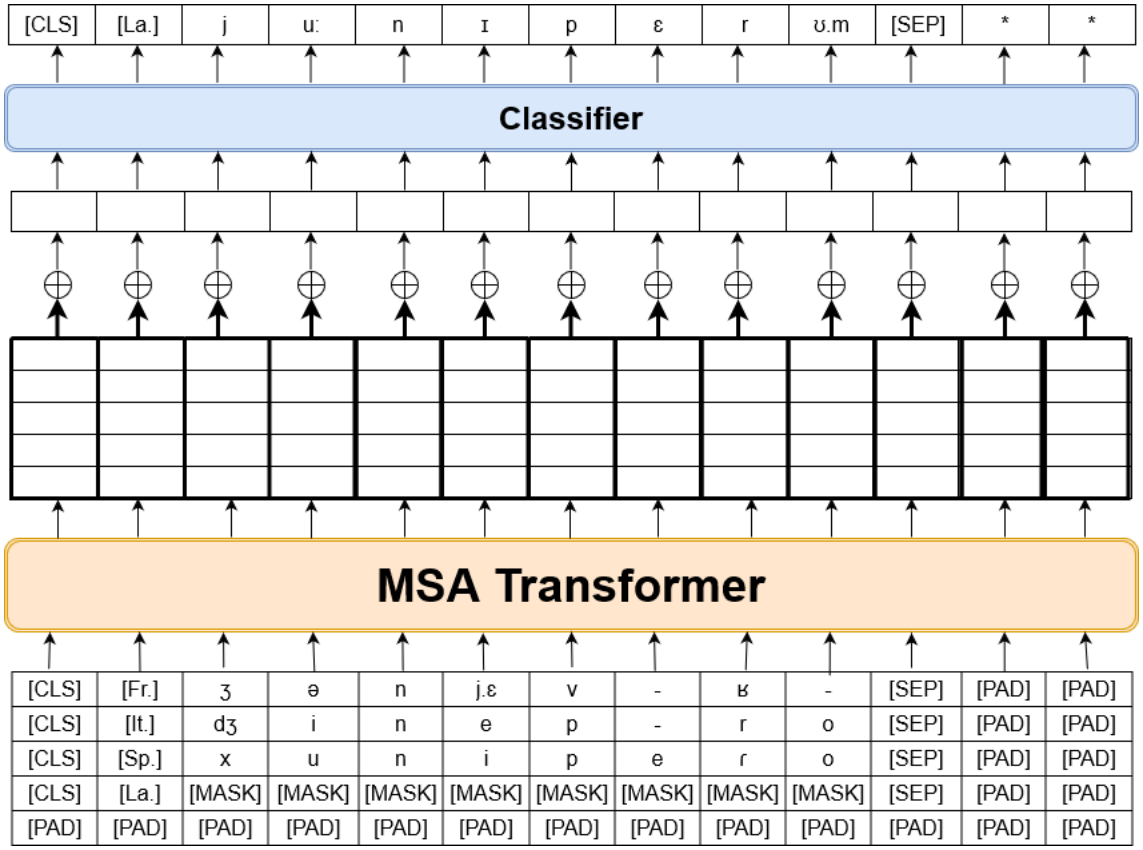


Figure 4.2: Cognate Transformer architecture: an input instance is passed into an MSA transformer, where the resultant embeddings are summed and normalized along columns, which are then finally passed into a classifier.

Meloni et al. (2021) incorporate the information regarding the language of a word through a language embedding concatenated to the character/token embedding. We instead treat *language information* as a separate token attached to the beginning of the phoneme sequence. Use of language embeddings with transformer based models was initially present in the multi-language model XLM (Conneau and Lample, 2019). It was however discontinued in the later versions(Conneau et al., 2020). We similarly have decided to remove the language embedding and instead use a special token denoting language as it is less complex in implementation. Other special tokens used include the usual [CLS] to mark the beginning, [SEP] to mark the ending of a word, [PAD] for padding, and [MASK] to replace '?' in the unknown word (see Table 4.2) or the word to be predicted. Thus, the input batch padded appropriately is passed on to the MSA Transformer.

The normal output of an MSA Transformer is a 2D array of embeddings per instance. To this, we add an addition layer that sums over columns to give a 1D array of embeddings

per instance as output. In other words, if the overall dimensions of the MSA transformer output were (`batch_size` × `num_languages` × `msa_length` × `hidden_size`) then, for our case, the final dimensions after summing up along columns are (`batch_size` × `msa_length` × `hidden_size`). To this, we add a normalizer layer followed by a classifier, i.e., a linear layer followed by cross-entropy loss. This is illustrated in Figure 4.2.

4.3.5 Pre-training

The described model can support pre-training in a form similar to masked language modeling where a word from a cognate set is entirely masked but the language token remains unmasked corresponding to the language that is to be predicted. In other words, *pre-training* is the same as training for cognate prediction task. For proto-language reconstruction, however, pre-training can be done. As a result, we pre-train Cognate Transfomer on the data of the cognate reflex prediction task. It is further *fine-tuned* on the proto-language reconstruction task.

We have used the publicly available implementation of MSA transformer by the authors[2], on top of which we added the layers required for the Cognate Transformer architecture. We have used tokenization, training, and related modules from HuggingFace library (Wolf et al., 2020). The entire code is made publicly available[3].

4.4 Experimental Setup

4.4.1 Datasets

We use the SIGTYP 2022 dataset (List et al., 2022b) for the cognate reflex prediction task. It consists of two different subsets, namely, training and surprise, i.e., evaluation data from several language families. The statistics for this dataset is provided in Table 4.3. Surprise data was divided into different test proportions (size of test data to that of training) of 0.1, 0.2, 0.3, 0.4, and 0.5 for evaluation. Among these, we only report for the test proportions

[2]https://github.com/facebookresearch/esm
[3]https://github.com/mahesh-ak/CognateTransformer

Family	Languages	Words	Cognates
Training data			
Tshanglic	8	2063	403
Bai	9	5773	969
Sino-Tibetan	7	1426	248
Sui	16	10139	1048
Uto-Aztecan	9	771	118
Afro-Asiatic	19	2583	340
Dogon	16	4405	971
Japonic	10	1802	278
Indo-European	4	1320	512
Burmish	7	2501	576
		32783	**5463**
Surprise data			
Atlantic-Congo	10	1218	388
Hui	19	9750	518
Chapacuran	10	939	187
Western Kho-Bwa	8	5214	915
Berta	4	600	204
Palaung	16	1911	196
Burmish	9	2202	467
Indo-European	5	565	212
Karen	8	2363	379
Bai	10	4356	658
		29118	**4124**

Table 4.3: Dataset for reflex prediction task

0.1, 0.3, and 0.5. The division into proportions is across different cognate sets and not within each. In other words, a cognate set in entirety would fall into either train or test data.

For the proto-language reconstruction task, the dataset provided by List et al. (2022a) is used. It consists of data from 6 language families, namely, Bai, Burmish, Karen, Lalo, Purus, and Romance whose statistics are listed in Table 4.4. This is divided into test proportion 0.1 by List et al. (2022a). We further test for proportions 0.5 and 0.8, where the latter proportion reflects a low-resource setting. For pre-training the Cognate Transformer for this task, we use the entire training data of both the tasks with words from proto-languages removed.

Family	Languages	Words	Cognates
Bai	10	3866	459
Burmish	9	1711	269
Karen	11	3231	365
Lalo (Yi)	8	7815	1251
Purus	4	693	199
Romance	6	18806	4147
		36122	**6690**

Table 4.4: Dataset for Proto-language reconstruction task

4.4.2 Model Hyperparameters

We have tested two variations of the proposed Cognate Transformer architecture, namely *CogTran-tiny* and *CogTran-small*. CogTran-tiny has hidden size 128, intermediate size 256, 2 attention heads, and 2 layers with overall 1 million parameters. CogTran-small has hidden size 256, intermediate size 512, 4 attention heads, and 4 layers with overall 4.4 million parameters. Both models have a vocabulary size of about 2,300.

For pre-training, only CogTran-small is used, since it consistently outperforms CogTran-tiny. The training is carried out with 48 epochs for pre-training, with 9 epochs for finetuning in the proto-language reconstruction task, 24 epochs for non-pre-trained in the same task, and 32 epochs for cognate-reflex prediction task, using Adam optimizer with weight decay (Loshchilov and Hutter, 2017) as implemented in HuggingFace transformers library (Wolf et al., 2020) with learning rate 1e-3 and batch size of 64. For finetuning the pre-trained model, the batch size is 48.

4.4.3 Evaluation

We use the metrics *average edit distance (ED)*, *average normalized edit distance (NED)*, and *B-Cubed F1 score (BC)* following List et al. (2022a) for evaluating the models. Edit distance is the well-known Levenshtein distance (Levenshtein, 1965), both with or without normalization by the lengths of the source and target strings being compared. B-Cubed F1 score (Amigó et al., 2009) was applied to phoneme sequences by List (2019b), where similarity is measured between aligned predicted and gold sequences. B-Cubed F1 score

measures the similarity in the structures and, hence, in the presence of systematic errors, carries less penalty than edit distance (See Appendix C for details on BCubed metrics. As (normalized) edit distance is a distance measure, the lower the distance, the better the model. On the other hand, for B-Cubed F1 it is opposite, i.e., the higher the score, the better the model. We import the metric functions from the `LingRex` package (List and Forkel, 2022).

4.4.4 Methods for Comparison

The results of the cognate reflex prediction task are compared directly against those of the top performing model in the SIGTYP 2022 task – Kirov et al. (2022). Here, direct comparison between the models is possible since the datasets including the test divisions are the same.

However, for the proto-language reconstruction task, the previous state-of-the-art model (Meloni et al., 2021) reports only on the Romance dataset with test proportion 0.12 and the baseline SVM model (List et al., 2022a) with additional features such as position, prosodic structure, etc., marked as SVM+PosStr is tested only with test proportion 0.1. However, the code is openly provided for the SVM-based model and, hence, results were generated for other test proportions 0.5 and 0.8 as well.

To compare the results of proto-language reconstruction with the NMT model given by Meloni et al. (2021) for which the code is not publicly available, we build a best-effort appropriate model identical to the one described there with 128 units Bidirectional GRU encoder followed by same sized GRU decoder followed by attention and linear layer with dimension 256 followed by a classifier. The input is encoded as a 96-dimensional embedding for each token concatenated with 32-dimensional language embedding. The training parameters are the same as previously stated in §4.4.2 except that the number of epochs trained is 32 and the batch size is 16. For the Romance data part, the results obtained are ED 1.287 and NED 0.157 whereas those reported by Meloni et al. (2021) for Romance data (IPA) with almost similar test proportion (0.12) are ED 1.331 and NED 0.119. Thus, the edit distances match whereas normalized ones do not. We speculate that

Test proportion	Method	ED↓	NED↓	BC↑
0.1	CogTran-tiny	1.0901	0.2997	0.7521
	CogTran-small	**0.8966**	**0.2421**	**0.7823**
	Mockingbird - Inpaint (Kirov et al., 2022)	0.9201	0.2431	0.7673
0.3	CogTran-tiny	1.3223	0.3497	0.6612
	CogTran-small	**1.1235**	0.2919	**0.6954**
	Mockingbird - Inpaint (Kirov et al., 2022)	1.1762	**0.2899**	0.6717
0.5	CogTran-tiny	1.4521	0.3873	0.6257
	CogTran-small	**1.2786**	**0.3332**	**0.6477**
	Mockingbird - Inpaint (Kirov et al., 2022)	1.4170	0.3518	0.6050

Table 4.5: Cognate reflex prediction results.

the NED reported by Meloni et al. (2021) could be erroneous due to possible inclusion of delimiter while calculating the length of the strings, since by (mis)considering delimiters, we obtain a similar NED 0.121 for the model we train. This can be confirmed by observing the ED-to-NED proportions of the corresponding scores obtained by the SVM-based model for the Romance dataset: ED 1.579 and NED 0.190, which we generate using the code made available by List et al. (2022a). Alternatively, the disparity in NED could also be attributed to differences in the sizes of the dataset used for training. However, it is unclear how agreement in ED score could have been then possible. Due to absence of both appropriate model and data, we assume that the NMT model we have built is a good reproduction of that built by Meloni et al. (2021).

All models compared in the proto-language reconstruction task are 10-fold cross-validated.

4.5 Results

In this section, we present and discuss in detail the results of our Cognate Transformer and other state-of-the-art models on the two tasks.

4.5.1 Cognate Reflex Prediction

The results of the cognate reflex prediction task are summarized in Table 4.5. The edit distance (ED), normalized edit distance (NED), and B-Cubed F1 (BC) scores are provided

Test proportion	Method	ED↓	NED↓	BC↑
0.1	CogTran-tiny	0.8081	0.1760	0.7946
	CogTran-small	0.7772	0.1683	0.7968
	CogTran-small Pretrained	**0.7459**	**0.1595**	**0.8081**
	SVM + PosStr (List et al., 2022a)	0.7612	0.1633	0.8080
	NMT GRU + Attn. (Meloni et al., 2021)	1.0296	0.1909	0.7560
0.5	CogTran-tiny	0.9013	0.1966	0.7279
	CogTran-small	0.8750	0.1899	0.7330
	CogTran-small Pretrained	**0.8177**	**0.1760**	**0.7534**
	SVM + PosStr (List et al., 2022a)	0.8455	0.1839	0.7425
	NMT GRU + Attn. (Meloni et al., 2021)	1.2585	0.2362	0.6733
0.8	CogTran-tiny	1.1043	0.2455	0.6781
	CogTran-small	1.0697	0.2359	0.6817
	CogTran-small Pretrained	**0.9754**	**0.2142**	**0.7132**
	SVM + PosStr (List et al., 2022a)	1.0630	0.2391	0.6800
	NMT GRU + Attn. (Meloni et al., 2021)	1.8640	0.3546	0.5538

Table 4.6: Proto-language reconstruction results.

Family \Test prop.	0.1	0.3	0.5
Atlantic-Congo	0.8192	0.7245	0.7027
Hui	0.7933	0.7418	0.7143
Chapacuran	0.6624	0.5847	0.5335
Western Kho-Bwa	0.8572	0.8065	0.7161
Berta	0.7681	0.6758	0.6197
Palaung	0.8815	0.7401	0.6863
Burmish	0.6531	0.5931	0.5261
Indo-European	0.5012	0.3915	0.3637
Karen	0.8869	0.7914	0.7504
Bai	0.8047	0.7029	0.6444

Table 4.7: Family wise B-Cubed F scores for model CogTran-small against test proportions on reflex prediction task

for Cognate Transformer across the test proportions 0.1, 0.3, and 0.5 along with the best performing model of the SIGTYP 2022 (List et al., 2022b) task, namely, the CNN inpainting (Kirov et al., 2022). CogTran-small consistently outperforms the previous best models across all test proportions. In particular, the difference in scores between Cognate transformer and the CNN inpainting model becomes prominent with increasing test proportion. Hence, it can be concluded here that Cognate Transformer is more robust than other models. The language family wise results for the best performing model, CogTran-small, are provided in Table 4.7.

Family \Test prop.	0.1	0.5	0.8
Romance	0.7765	0.7570	0.7353
Bai	0.7465	0.7108	0.6748
Burmish	0.8426	0.7250	0.6460
Karen	0.8666	0.7845	0.7373
Lalo	0.7221	0.6769	0.6416
Purus	0.8941	0.8662	0.8440

Table 4.8: Family wise B-Cubed F scores for model CogTran-small Pretrained against test proportions on proto-language reconstruction task

Test prop.	ED	NED	BCF
0.1	0.065	0.015	0.018
0.5	0.028	0.006	0.008
0.8	0.027	0.006	0.007

Table 4.9: Standard Deviations for model CogTran-small Pretrained

4.5.2 Proto-Language Reconstruction

The results of the proto-language reconstruction task are summarized in Table 4.6 with the same evaluation metrics along with comparisons with other previously high performing models, namely, SVM with extra features by List et al. (2022a) and NMT (GRU-attention) based by Meloni et al. (2021) for the test proportions 0.1, 0.5, and 0.8. Previously, there were no comparisons between SVM-based and NMT-based models. Here, we find that the SVM-based model performs consistently better than the NMT-based model. In other words, the GRU-Attention-based NMT model does not appear to scale well in harder situations, i.e., for higher test proportions when compared with the other models. While CogTran-small achieves results similar to the SVM-based models, pre-training makes a difference. The pre-trained Cognate transformer outperforms all the other models in all test proportions. Although the increase in the proportion 0.1 is not much significant, paired t-test between best performing model and the next best model i.e. CogTran-small Pretrained and SVM-based yield significance of $p < 0.01$ in low-resource proportions i.e. 0.5 and 0.8 . The language family wise results and standard deviations for the best performing model, CogTran-small Pretrained are provided respectively in Table 4.8 and Table 4.9. Note that SVM-based model was also part of SIGTYP 2022 (List et al., 2022b) where it lags well behind CNN inpainting model. Hence, cognate transformer generalizes

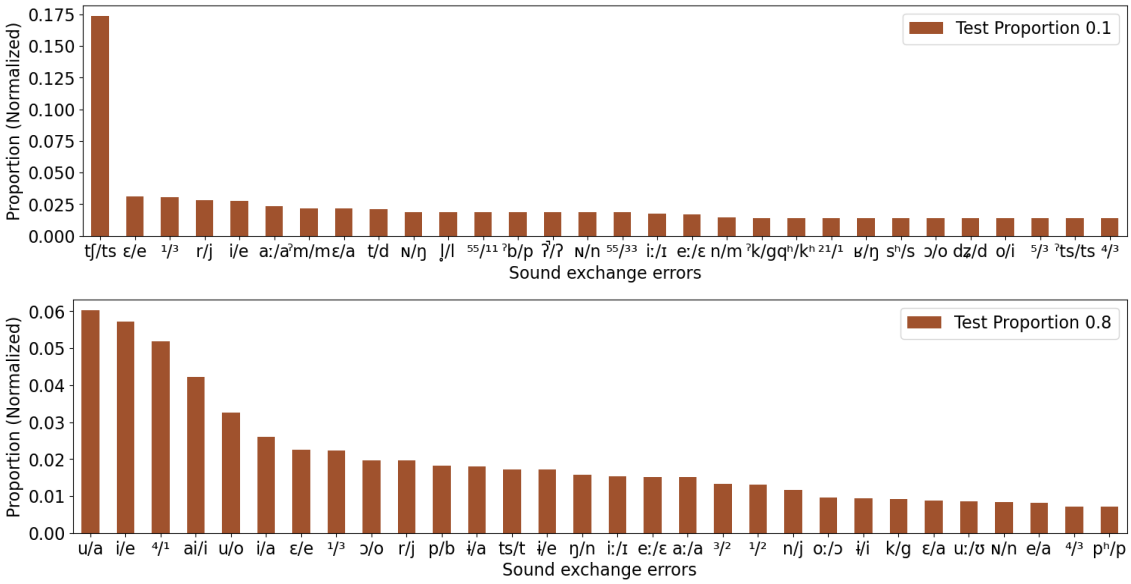


Figure 4.3: Top-30 most common sound exchange errors out of over 400 errors for pre-trained CogTran-small on proto-language reconstruction task with test proportions 0.1 (top) and 0.8 (bottom).

well across tasks hence gains from architecture are obvious.

4.5.3 Error Analysis

To analyze errors, we consider the pre-trained and finetuned CogTran-small on the proto-language reconstruction task for the easiest and hardest test proportions 0.1 and 0.8 over fixed data (without cross-validation). Figure 4.3 shows the 30 most common sound exchange errors by the models. An example of sound exchange error, u/a means either 'a' is predicted in place of 'u' or vice versa. To make this plot, we first gather the frequencies of sound exchanges for the various language families in data by comparing the aligned predicted and gold reconstructions. These frequencies are normalized for each proto-language or language family and finally combined and normalized again. Normalization at the language family level is important since few language families show more tendencies for certain types of errors than others. Since data is not equally available from all families, a language family with more data influences the outcome. For example, among the datasets used for the task, the Romance dataset comprises half of them. We observe that Romance data shows more vowel-length-related errors as also observed by Meloni et al. (2021) and, thus, proportion of such errors is inflated. Hence, normalization is carried out at the

language family level to prevent such biases. We normalize per family by dividing the frequency of a particular error type in a family by the total number of errors in that family. Normalized frequencies thus obtained per error type per family are combined by adding up across families and then normalized again.

The most frequent sound exchange errors are plotted in Figure 4.3 which make up respectively, for test proportions 0.1 and 0.8, about 71% and 60% of total such errors. One can observe from the plot that the most common vowel errors are the exchange of short vowels /u/ and /i/ with a neutral vowel /a/, vowel raising-lowering, i.e., exchange of /i/ $\sim$ /e/, /u/ $\sim$ /o/, diphthong-monophthong exchanges /ai/ $\sim$ /i/, tense-laxed exchanges, i.e., /ɛ/ $\sim$ /e/ and /ɔ/ $\sim$ /o/. Vowel length confusions, i.e., /iː/ $\sim$ /ɪ/, /eː/ $\sim$ /e/, /aː/ $\sim$ /a/, /oː/ $\sim$ /ɔ/, /uː/ $\sim$ /ʊ/ also make up a significant portion. Overall, vowel/consonant length errors make up to about 10% sound exchange errors each in both cases. Among consonant errors, one can observe voiced-unvoiced or glottalized-unglottalized consonant exchanges like /p/ $\sim$ /b/, /ˀk/ $\sim$ /g/, aspiration errors, i.e., /pʰ/ $\sim$ /p/, /tʰ/ $\sim$ /t/, change of place of articulation like /ŋ/ $\sim$ /n/, /s/ $\sim$ /h/, etc. Tone exchange errors like /¹/ $\sim$ /³/ also exist which contribute to about 10% in each of the cases. Affricatives exchange error /tʃ/ $\sim$ /ts/ appears prominently in the case of test proportion 0.1. Overall, these are the most general kinds of errors; however, exact types of errors are observed to be dependent on the language family. Hence, although most general ones are universally observed, significant differences can be expected based on the particular datasets or perhaps due to learning of different sound correspondences in different runs by the network.

4.5.4 Zero-shot Attempt

Previously, we discussed the results of proto-language reconstruction for various test proportions. Among these, the highest proportion considered, i.e., 0.8, can be thought of as a *few-shot* learning case, since for some of the language families like Purus and Burmish, the number of training instances, i.e., cognate sets is less than 50. We next consider the pre-trained model for the same task without any finetuning; in other words, we consider the *zero-shot* case. The scores achieved by such a model are 2.6477 ED, 0.5758 NED,

and 0.5499 BC, which means that more than 40% of a word on average in generated reconstructions are correct. An example input instance and its corresponding output and gold data from the Romance dataset:

Input: [Latin] ?,
[French] p ɛ ʁ s p i k y i t e,
[Italian] p e r s p i k u i t a
Output: [French] p e r s p i k y i t a
Gold: [Latin] p ɛ r s p ɪ k ʊ ɪ t aː.t.ɛ.m

In the above example, the output language token is incorrect. Since the proto-languages (in this case, Latin) have been excluded entirely in pre-training, this can be expected. One can also observe that the output word completely agrees with neither Italian nor French, although the inclination is more toward the former (with a ED of 1). A similar observation was made by Meloni et al. (2021) where the network attended most to Italian since it is conservative than most other Romance languages.

4.5.5 Learned Sound Changes

Here, we consider the finetuned pre-trained model on the proto-language reconstruction task to observe the learned sound changes by the network in the hardest scenario, i.e., with test proportion 0.8. The following example reveals an instance where palatalization appearing in Romance languages is correctly reconstructed to a non-palatal consonant:

Input: [Latin] ?,
[French] s j ɛ,
[Spanish] ɵ j e
Output: [Latin] k ɛ

We now consider *metathesis*, a non-trivial complex sound change where positions of phonemes are interchanged. The following example is from the training set which the network learns correctly and demonstrates the metathesis *-bil-* > *-ble-*.

Input: [Latin] ?,

[French] ɛ̃ p ɛ ʁ s ɛ p t i b l,

[Spanish] i m p e ɾ θ e pː t i β l e

Output: [Latin] ɪ m p ɛ r k ɛ p t ɪ b ɪ l ɛ m

Following is an example from the test set where the model confuses a complex metathesis pattern occurring in Hispano-Romance, *-bil-* > *-lb-*.

Input: [Latin] ?,

[Spanish] s i l β a ɾ,

[French] s y b l e,

[Portuguese] s i l v a ɪ

Output: [Latin] s y b l w aː r ɛ

Gold: [Latin] s iː b ɪ l aː r ɛ

Even the model finetuned on test proportion 0.1 does not get this example correct. Its output is

Output: [Latin] s y b l ɔ aː r ɛ

Hence, metathesis can be seen as a hard sound change to be learned by this model. This is not surprising since metathesis or site exchange does not naturally fit into the sequence alignment approach which fundamentally only models insertions and deletions at any site. Thus, it is worthwhile to investigate more on this aspect by training the network on language families that exhibit systematic metathesis to understand its behavior.

4.6 Summary

In this chapter, we adapted MSA transformer for two phonological reconstruction tasks, namely, cognate reflex prediction and proto-language reconstruction. Our novel architecture, called Cognate Transformer, performs either comparable to or better than the previous methods across various test-train proportions consistently. Specifically, the pre-trained model outperforms the previous methods by a significant margin even at high test-train proportions, i.e., with very less trainable data reflecting a more realistic scenario.

To the best of our knowledge, this work demonstrates the utility of transfer learning

when applied to historical linguistics for the first time. In this chapter, the data is in IPA representation, but this is not necessary as long as words can be aligned with properly defined sound classes in the respective orthographic representations. This may not be an easy task when the languages are phonologically diverse. A fixed transliteration scheme may be used in such cases if feasible. Thus, relaxing the IPA input constraint can increase the amount of trainable data and pre-training with more data would most likely improve the performance of not only the problem of automated phonological reconstruction but can be demonstrated in the future for an important related task, namely automated cognate word detection. Further, more standard ways of pre-training such as masking only a couple of tokens across all languages instead of a complete word of a single language can be adapted in future.

Limitations

In the task of proto-language reconstruction, it can be seen from the results (Table 4.6) that CogTran-small i.e. the plain Cognate Transformer model without pre-training slightly underperforms the SVM-based model at low test proportions. Only the pre-trained model performs well in this scenario.

Further, it has already been mentioned in §4.5.5 that metathesis sound change is not being captured correctly by the network which requires further investigation. Overall, very few languages and language families are included in the data used. Thus, it is desirable to create such datasets for other languages with at least cognacy information to improve the unsupervised training firstly, which can be then employed in supervised training successfully with fewer training examples.

Chapter 5

Cognate Detection as a Link Prediction Task with Cognate Transformer

Identification of cognates across related languages, as described in Chapter 2, is one of the primary problems in historical linguistics. Automated cognate identification is helpful for several downstream tasks including identifying sound correspondences, proto-language reconstruction, phylogenetic classification, etc. Previous state-of-the-art methods for cognate identification are mostly based on distributions of phonemes computed across multilingual wordlists and make little use of the cognacy labels that define links among cognate clusters. In this chapter, we present a transformer-based architecture inspired by computational biology for the task of automated cognate detection. Beyond a certain amount of supervision, this method performs better than the existing methods, and shows steady improvement with further increase in supervision, thereby proving the efficacy of utilizing the labeled information. We also demonstrate that accepting multiple sequence alignments as input and having an end-to-end architecture with link prediction head saves much computation time while simultaneously yielding superior performance. This chapter is based on Akavarapu and Bhattacharya (2024a).

5.1 Introduction

Words in genetically related languages with same descendance from a common ancestral language are termed as *cognates*. For example, Sanskrit *bhava* and English *be* are cognates reconstructed as **b^{h}ewH-* in ancestral Proto-Indo-European. Within historical linguistics, assembling potential cognates forms an essential step in the comparative method to proceed to further stages such as formulation of sound laws, reconstruction of proto-language, phylogenetic reconstruction, etc. (Campbell, 2013). Cognate identification has been traditionally carried out by tedious manual cross-comparisons of lexica across several concepts or meanings; this often requires sufficient linguistic expertise in the languages that are being compared. *Automated cognate detection* attempts to alleviate manual labor and, thus, assists a historical linguist to quickly produce high-quality etymologies required for downstream tasks.

Over the past decade, several methods for automated cognate detection, mostly using sequence alignment and other techniques inspired by bioinformatics and evolutionary biology (List et al., 2017), have appeared. The best-performing methods primarily depend on similarity scores computed from distributions of phonemes in multilingual wordlists (Rama and List, 2019) and make little or no use of the cognacy labels except for a clustering task at the end. In this chapter, we advocate for a supervised learning scenario that utilizes the labeled information to the fullest. We demonstrate that such a scenario combined with the representational power of an appropriate deep neural network architecture can outperform previous methods above a certain amount of supervision. We also demonstrate that such a model is also capable of transfer learning. In other words, once trained on some data, it can perform well on any dataset unseen so far with little additional supervision.

The typical procedure followed by the state-of-the-art methods for this problem is as follows. In each language family, attested words from all languages that have the same meaning, i.e., *concept*, are clustered based on the pairwise similarity measures computed by the respective procedure. We propose a different approach where instead of clustering based on pairwise similarity we directly take input a *multiple sequence alignment (MSA)*

of words of the same concept and predict linkage via an end-to-end architecture. This approach proves to be much better in performance and much faster than clustering from independent pairwise similarity measures.

Many of the algorithms in computational historical linguistics are heavily drawn or inspired by computational biology. Continuing the trend, we adopt Cognate Transformer (Akavarapu and Bhattacharya, 2023a), which yielded state-of-the-art performance in automated phonological reconstruction task, as the base architecture. Cognate Transformer was adapted from MSA Transformer (Rao et al., 2021), a protein language model that excels in contact predictions. We additionally append to this architecture layers consisting of triangular multiplication and triangular attention modules inspired by Alphafold2 (Jumper et al., 2021), the state-of-the-art protein structure predictor, where the modules roughly capture triangle inequalities among the distances between amino acid residues. For our task, we applied these modules for capturing transitivity property among linkages in cognate clusters. We find that the addition of this particular module has a significant share in the performance of the overall architecture.

Our key contributions are as follows:

1. Firstly, we propose a supervised method for automated cognate detection that outperforms existing methods with sufficient supervision with likely improvement on further supervision, thus utilizing the labeled data much more efficiently than previous models while also demonstrating few-concept (akin to few-shot) learning.

2. Secondly, our method consists of an end-to-end architecture that avoids independent pairwise computations by accepting MSA as input and directly predicting cluster linkages, which proves to be more efficient in terms of both performance and time than a pairwise approach.

3. Thirdly, we incorporate into the architecture of Cognate Transformer additional modules to capture transitivity property among cognate cluster linkages which has a positive effect on overall performance.

The rest of the chapter is organized as follows. Related work is mentioned in §5.2. The

problem statement is elaborated in §5.3. The methodology is described in §5.4. The details of the experimental setup including the datasets used, previous baselines, and evaluation measures are described in §5.5. The results of experiments and ablation studies along with error analyses and discussions are given in §5.6. Finally, the article is concluded in §5.7.

5.2 Related Work

Computational historical linguistics is a young field that emerged over the past two decades. Notable works that lead to significant progress in automatic cognate detection are as follows. Consonant Class Method of Turchin et al. (2010) deems two words as cognate if the first two consonants fall under the same consonant class. In Sound-Class-based phonetic alignment (SCA) of List (2010), pairwise phoneme sequences are aligned and scored for similarity using sound classes that extend consonant classes. LexStat (List, 2012a) aligned and scored pairwise sequences using language phonemic-specific distributions combined with SCA-based scores. The pairwise similarities thus obtained are clustered using UPGMA (Sokal and Michener, 1958). The previous state-of-the-art results are attributed to LexStat combined with Infomap clustering (List et al., 2017). Equivalent performance was also reported in Rama (2018) using Chinese Restaurant Clustering. An expectation-maximization method over pairwise phonemic distributions is also found to yield similar performance (MacSween and Caines, 2020). Information-weighted similarity measure was proposed by (Dellert, 2018) which reported a slight increase in evaluation scores over LexStat, albeit tested only on one dataset.

Supervised algorithms include the Siamese-CNN-based model by Rama (2016) which performs binary classification on a given pair of words. Jäger et al. (2017) employ SVM on top of LexStat and Point-wise Mutual Information (PMI) measures that yield performance similar to that of LexStat-Infomap.

There exist several other works often performing supervised pairwise classification and incorporating multilingual language models such as those of Kanojia et al. (2020, 2021) and Nath et al. (2022). Despite brilliantly employing pre-trained multilingual language models, these cannot be applied for ancient languages like Ancient Greek, Gothic, etc.,

or highly low-resource and endangered languages like those of the Americas where one does find wordlists of sufficient size but not enough text to pre-train language models for sake of performing historical linguistic tasks computationally. Another related task is that of cognate and derivate detection (Rani et al., 2023), which is essentially a word-pair classification task. These tasks have a slightly different setup than the problem at hand since the clustering step is not involved.

Cognate Transformer (Akavarapu and Bhattacharya, 2023a), described in the previous chapter, that achieves the best performance on phonological reconstruction tasks employs a transformer-like architecture with row-wise and column-wise attentions to efficiently operate over MSAs. This model was adapted from an evolutionary biological model called MSA Transformer (Rao et al., 2021) which acts on protein sequences. Vanilla Transformer architecture was also used in Kim et al. (2023) for proto-language reconstruction. Although we employ Cognate Transformer, it should be well noted that the problem we are addressing is that of cognate detection which is quite different from that of proto-language reconstruction. The aforementioned transformer-based models address the latter problem.

5.3 Automated Cognate Detection

The automated cognate detection problem statement is described here as follows. The gold data for a language family F, comprising of related languages $L_1, L_2, \ldots \in F$, consists of words over several concepts, i.e., meanings, say $M_1, M_2, \ldots$, etc. Each word is a sequence of phonemes. For each concept M_m, there are words W_i^m for several languages L_i in that family, where W_i^m is a word of a language L_i in concept M_m. Words in each concept are associated with labels say $c_i^m \in \mathbb{N}$ which indicate the cluster to which they belong. A single such cluster of words is called a *cognate set*. We also define links $l_{ij}^k \in \{0, 1\}$ between languages L_i and L_j for a concept M_m which indicate if the corresponding words

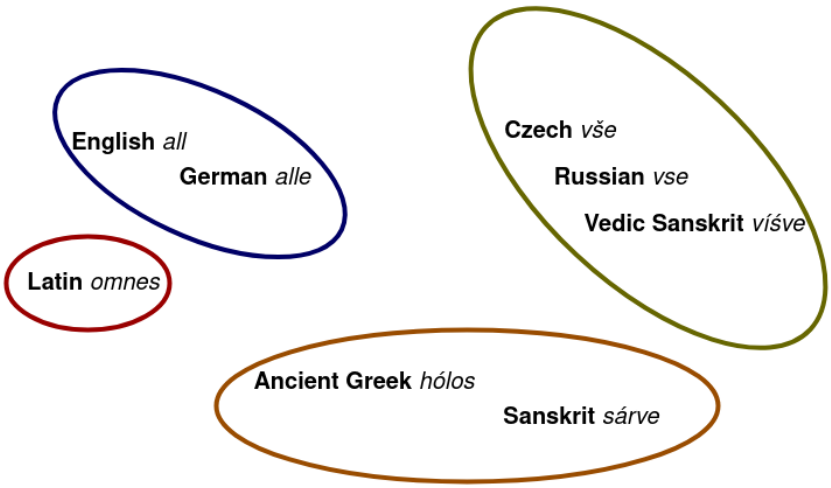


Figure 5.1: Examples of cognate clusters for the concept 'all' in few Indo-European languages.

are cognates i.e., have the same cluster label. In other words,

$$l_{ij}^{m} = \begin{cases} 1 & \text{if } c_i^m = c_j^m \\ 0 & \text{if } c_i^m \neq c_j^m \end{cases} \tag{5.1}$$

The goal of automated cognate detection is to correctly cluster a given set of words that mean a single concept in a language family. In a supervised setting, the aim is to predict the linkages correctly.

For an illustration of the overall problem, consider the Indo-European language family and the concept of 'all'. The attested lexica in the member languages are Sanskrit *sárve* (Vedic *víśve*), Greek (Ancient) *hólos*, Latin *omnes*, German *alle*, English *all*, Russian *vse*, Czech *vše*, etc. Among these Vedic *víśve*, Russian *vse*, Czech *vše* form a cluster, i.e., a cognate set while Sanskrit *sárve* and Greek *hóla* form another cognate set. Similarly, English and German word forms form another cognate set. The input data is present in IPA transcription format. Roman transliterated forms are presented here only for demonstration. See Figure 5.1 as an illustration of this example.

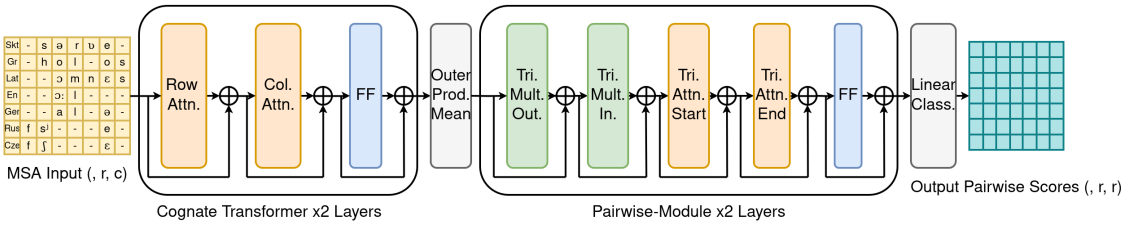


Figure 5.2: Architecture of Cognate Transformer with Triangular Multiplication and Attention modules

Skt.	-	s	ə	r	ʊ	e	-
Gr.	-	h	o	l	-	o	s
Lat.	-	-	ɔ	m	n	ɛ	s
En.	-	-	ɔː	l	-	-	-
Ger.	-	-	a	l	-	ə	-
Rus.	f	s^j	-	-	-	e	-
Cze.	f	ʃ	-	-	-	ɛ	-

Table 5.1: Example of a Multiple Sequence Alignment (MSA) of phoneme sequences

5.4 Methodology

The overall workflow is described as follows. Given some words from different languages for a concept in a language family, the words are first aligned (§5.4.1), then converted into tokens and passed into the cognate transformer (§5.4.2), whose outputs are converted into pairwise (along language axis) representations by outer product mean module (§5.4.3), which are then passed into the layers of pairwise module (§5.4.4) whose outputs are classified into two labels 0 or 1 indicating the pairwise linkage among the languages (§5.4.5). Since the linkage information is known in the form of cognacy labels, the architecture described can be thus trained end-to-end. The overall architecture is illustrated in Figure 5.2.

5.4.1 MSA input

The input words for a concept are aligned together using the SCA method (List, 2010), where initial pairwise alignments are carried out by using Needleman and Wunsch (1970) with weights based on sound classes which are further progressively merged guided by a UPGMA (Sokal and Michener, 1958) tree based on pairwise distances. Progressive

alignment is a widely used method for multiple sequence alignment which forms the basis of popular programs such as ClustalW (Thompson et al., 2003). We use the implementation available in LingPy (List and Forkel, 2021).

The resultant MSA, present in IPA (see Table 5.1), is converted into ASJP (Brown et al., 2008) representation, a phonemic representation scheme that compacts IPA symbols resulting in lesser vocabulary size. Note that each token in an MSA need not be a single phoneme. In the SCA method, consecutive vowels are combined into one token. Language information is passed as the initial token in each row following Akavarapu and Bhattacharya (2023a). The resultant tokens are mapped to their respective token numbers and padded according to the batch. Thus, a typical input to Cognate Transformer lies in $\mathbb{N}^{b \times r \times c}$ where b is the batch size, r is the maximum number of rows, i.e., the number of words for that batch, and c is the maximum sequence length in the batch. From here, we ignore the batch dimension and simply consider the input to lie in $\mathbb{N}^{r \times c}$

5.4.2 Cognate Transformer

Cognate Transformer (Akavarapu and Bhattacharya, 2023a) handles two-dimensional input employing separate row and column attentions (see Figure 5.2). The input and output have the same dimensions. In other words,

$$\text{CogTran} : \mathbb{N}^{r \times c} \rightarrow \mathbb{R}^{r \times c \times d} \tag{5.2}$$

where d is the hidden size. The outputs of CogTran are converted into pairwise format by the outer product mean module.

5.4.3 Outer Product Mean

In this module, as the name suggests, the outer product is computed along each column, across all rows, and then the mean of outer products is computed across all columns. The

transformation to the dimensions are

$$\text{OutProdMean} : \mathbb{R}^{r \times c \times d} \rightarrow \mathbb{R}^{r \times r \times d} \tag{5.3}$$

The intuition is that the multiplication of a pair of transformed embeddings of two tokens in a single position (i.e., column) of two different words (i.e., rows) should roughly indicate the similarity or distance between the two words in that particular position. The mean operation should produce a mean of such similarities across all positions for a pair of words. Hence, the final matrix would represent a pairwise similarity matrix across the words in an MSA.

This module is identical to the one in AlphaFold2 (Jumper et al., 2021) except that the role of rows and columns is interchanged. In other words, in AlphaFold2, the outputs are pairwise representations of amino-acid-residues (along columns) while in our case the outputs are pairwise representations of words (along rows).

5.4.4 Pairwise Module

The pairwise module in AlphaFold2, which consists of triangle multiplication and triangle attention updates via both incoming and outgoing edges, is a differentiable workflow to capture triangle inequalities that the distances between amino acid residues should satisfy (Jumper et al., 2021). In our case, we demand that the link predictions (see §5.3 for definition) satisfy the transitivity property which can be translated into the following condition

$$l^m_{ik} \cdot l^m_{jk} = l^m_{ij} \text{ if } l^m_{ik} + l^m_{jk} \neq 0 \tag{5.4}$$

for languages L_i, L_j and L_k in a family F for concept M_m. The triangle multiplication update follows a similar equation but without constraint and, hence, is apt for the problem at hand. Combining the updates for both incoming ($i \rightarrow j$) and outgoing edges ($j \rightarrow i$) ensures the symmetry required for pairwise similarities. The pairwise module does not

alter the dimensions of the input, i.e.,

$$\text{PairwiseMod} : \mathbb{R}^{r \times r \times d} \rightarrow \mathbb{R}^{r \times r \times d} \tag{5.5}$$

In AlphaFold2, this module along with the MSA module is embedded within the Evoformer module. As of now, it is unclear if such embedding would improve the performance. For this problem, we stack the modules as illustrated in Figure 5.2 for the sake of simplicity and easier ablation tests.

5.4.5 Classifier and Clustering

The outputs of the pairwise module are passed through a linear layer outputting values for two classes $\{0, 1\}$ indicating linkage. Hence, the classifier layer's transformation is summarized as:

$$\text{Classifier} : \mathbb{R}^{r \times r \times d} \rightarrow \mathbb{R}^{r \times r \times 2} \tag{5.6}$$

The softmax probabilities of the outputs p_{ij}^m for $P(l_{ij}^m = 1)$ determine the linkage probabilities. In addition to the above 'soft' constraints, the links are also expected to satisfy the symmetric property, i.e., $l_{ij}^m = l_{ji}^m$. While there are no special modules included to ensure this property, labels are considered symmetrically (as opposed to only lower/upper triangle) while computing the loss. During training, the network is trained with cross-entropy loss. During testing, UPGMA is run for each concept M_m with pairwise similarities as p_{ij}^m flat clustered at a threshold of 0.6, which is determined by a small (5%) held out validation set during training, to obtain the required clusters.

For additional information on the outer product mean and pairwise modules, see Appendix B.

Family	Meanings	Languages	Cognates	Words
Training data				
AN	210	20	2864	4358
BAI	110	9	285	1028
CHN	140	15	1189	2789
IE	207	20	1777	4393
JAP	200	10	460	1986
OU	110	21	242	2055
Total			**6817**	**16609**
Test data				
BAH	200	24	1055	4546
CHN	180	18	1231	3653
HU	139	14	855	1668
ROM	110	43	465	4853
TUJ	109	5	179	513
URA	173	7	870	1401
AN	210	45	3804	9267
AA	200	58	1872	11827
IE	208	42	2157	9854
PN	183	67	6634	12691
ST	110	64	1402	7074
Total			**19136**	**67347**

Table 5.2: Details of the datasets as obtained from Rama and List (2019) indicating the number of concepts, languages, cognate sets, and words.

5.5 Experimental Setup

In this section, the details of the experiments including datasets, implementation, evaluation metrics, baseline models, etc. are described.

5.5.1 Datasets

The dataset for both training and testing along with the train-test split is taken from Rama and List (2019) which was collected from various publicly available sources. It consists of data from various language families, namely, Austro-Asiatic (AA), Austronesian (AN), Bai (BAI), Bahnaric (BAH), Chinese (CHN), Huon (HU), Indo-European (IE), Japanese (JAP), Ob-Ugrian (OU), Pama-Nyungan (PN), Romance (ROM), Sino-Tibetan (ST), Tujia (TUJ), and Uralic (URA). The statistics of the data are provided in Table 5.2.

As is evident from the table, the original training size is disproportionately much lesser than the test size. Many language families in tests such as AA, PN, HU, etc. are completely absent in the training set. We also test the model on increased supervision by augmenting the training data with some proportion of test data. In particular, apart from the original train-test split, we also test by including 12.5% and 50% additional test concepts, i.e., approximately 20 and 100 additional test concepts respectively per language family. For both the proportions, data is divided into 5 random splits. Hence, the results reported for 12.5%+ and 50%+ proportions are five-fold cross-validated.

5.5.2 Implementation Details

The architecture we deploy has two Cognate Transformer layers and two layers of pairwise module (see Figure 5.2). In the Cognate Transformer, the number of attention heads is also 2. The maximum vocabulary size of the tokenizer is set to 768, while the maximum words and sequence length in an MSA are both set to 256. Both hidden size d and intermediate size, wherever there is projection, are 128. This amounts to a network of about a million parameters. The network was trained with a batch size b of 4 and tested with that of 2. Low batch size is due to the limitation of GPU memory (10 GB in our case) since MSAs combined in both the dimensions and the pairwise representation layers easily blow up the memory. The training was performed using AdamW optimizer (Loshchilov and Hutter, 2017) with learning rate 1e-3 as implemented by HuggingFace (Wolf et al., 2020). During testing, the pairwise softmax probabilities (similarities with 1 being the most similar) are used for flat clustering using UPGMA at a threshold of 0.6, arrived through held-out validation from the train set (5%). The total time taken for one run of train and test is less than 15 minutes on GPU. This is much smaller when compared to the models that operate on a pair of words at a time instead of on an MSA. The code is made publicly available[1].

[1] https://github.com/mahesh-ak/CogDetect

5.5.3 Evaluation Metrics

The outputs of the entire algorithm are clusters (see §5.4.5), i.e., every word gets a cluster label assigned which is to be compared with the gold cluster labels. The usual F1 score is not a proper measure since the assigned cluster label is not important; rather, members of the same cognate set must get assigned to the same cluster. Hence, the B-Cubed F1 score (Amigó et al., 2009) is the appropriate evaluation measure; it has been employed in the previous works for this problem as well (see Appendix C for details on BCubed metrics). We use the implementation available in LingPy (List and Forkel, 2021).

5.5.4 Baseline Models

LexStat-Infomap

We label the model defined so far as *CogTran2*. The foremost base model with which we compare the performance of CogTran2 is LexStat-Infomap (List et al., 2017) whose performance is more or less the state-of-the-art as discussed in §5.2. The original model employs 10,000 permutations between each language pair in a family to obtain language-specific distributions. Hence, this method requires significant test data to be known beforehand to preprocess. We call this model as *LexStInf10K*. This method takes more than 2 hours on a CPU to obtain results on one test set. Hence we also report for the model that has the number of runs as 1000, which we label as *LexStInf1K* which takes less than 15 CPU minutes. These are imported from LingPy (List and Forkel, 2021).

SCA

We also test on SCA-based model (List, 2010) where a pairwise distance depends on sound classes and alignment. Since it does not depend on any sort of computation such as language-specific distributions, this is the fastest method and, unlike LexStat-Infomap, can be run on any unseen data. We label this as *SCA*. For both LexStat-Informap and SCA, we use the flat cluster thresholds 0.6 and 0.45 respectively, as mentioned in List et al. (2017), since the training data is the same.

SVM

We also compare with the SVM-based model (Jäger et al., 2017), labeled as *SVM*, and the Siamese-CNN-based model (Rama, 2016) as these are supervised models. This model uses LexStat score and PMI scores as primary features and, hence, takes a long time to preprocess data, i.e., about 6 hours when each split is processed in parallel on a CPU when the number of permutations runs is 1000 (for LexStat similarity). Since this is a relatively much longer time, we do not increase the number of runs any further. SVM is trained on pairwise binary classification tasks which give pairwise cognacy probabilities for further clustering. We use publicly available code for this model[2].

Siamese CNN

From the proposed Siamese CNN architectures (Rama, 2016), we use the model mentioned as charCNN with language features that show good overall performance among the models that are proposed therein. We label this model as *CharCNN*. The network is trained on pairwise supervised binary classification tasks. The pairwise probabilities of the network are used further for clustering (UPGMA). CharCNN is implemented from scratch in PyTorch closely following the TensorFlow code that was made publicly available by the author [3].

Ablation Models

We also test on ablations, namely, without pairwise module which we call simply *CogTran*.

We also test by increasing the number of hidden layers to 4 of this same model which we label as *CogTranL4*.

Further, we test on a variant that does not use input MSA but rather only an alignment of a pair of words at a time akin to all other previous models but unlike CogTran2. In this model, pairwise binary classification is performed which gives probability scores for each pair of words in a concept. Further, clustering (UPGMA) is performed using these pairwise

[2]https://github.com/evolaemp/svmcc
[3]https://github.com/PhyloStar/SiameseConvNet/

scores. To be more specific, the input is an aligned word pair and the resultant output embeddings are summed before the binary classifier, while in Siamese-CNN (Rama, 2016), the absolute differences of embedding pairs are considered before the classifier layer. We note that summing should not be different since the network can always adjust the signs within embeddings themselves. We call this model *CogTranPair*. For these models, the link prediction is not part of the end-to-end architecture, unlike for the model we propose. As a result, the models are run separately on all possible pairs of words in a concept.

Data+%	Method	Language Families											Mean
		BAH	CHN	HU	ROM	TUJ	URA	AN	AA	IE	PN	ST	
+0%	SCA	.864	.793	.857	.873	.894	.909	.775	.760	.806	.709	.561	.800
	LexStInf10K	.894	.857	**.883**	.910	**.899**	**.913**	.840	**.773**	.826	**.845**	.592	**.839**
	LexStInf1K	.894	.855	.873	.912	.900	.907	.839	.759	.818	.820	**.595**	.834
	CharCNN	.759	.837	.876	.666	.845	.886	.698	.722	.725	.784	.473	.752
	SVM	.865	.845	.860	**.927**	**.899**	**.913**	**.845**	.734	.828	.782	.593	.826
	CogTran2	.854	**.864**	.857	.907	.893	.899	.786	.756	**.845**	.797	.572	.821
+12.5%	CharCNN	.830 (.010)	.847 (.006)	.873 (.010)	.896 (.007)	.892 (.015)	.895 (.006)	.777 (.007)	.752 (.007)	.825 (.008)	.786 (.002)	.535 (.025)	.810 (.002)
	SVM	.878 (.006)	.836 (.006)	.882 (.010)	**.934 (.007)**	**.919 (.005)**	**.914 (.006)**	**.840 (.003)**	.767 (.012)	.831 (.004)	.765 (.012)	.582 (.012)	.832 (.002)
	CogTran2	**.884 (.004)**	**.867 (.005)**	**.890 (.011)**	.907 (.013)	.913 (.015)	.904 (.006)	.810 (.005)	**.813 (.003)**	**.851 (.003)**	**.804 (.007)**	**.607 (.020)**	**.841 (.002)**
+50%	CharCNN	.876 (.011)	.854 (.007)	.880 (.005)	.914 (.012)	.899 (.018)	.904 (.012)	.795 (.008)	.784 (.005)	.840 (.006)	.785 (.011)	.563 (.011)	.827 (.003)
	SVM	.881 (.010)	.838 (.009)	.889 (.014)	**.935 (.010)**	**.927 (.012)**	**.914 (.009)**	**.840 (.010)**	.779 (.009)	.828 (.007)	.775 (.009)	.577 (.019)	.835 (.002)
	CogTran2	**.893 (.011)**	**.878 (.005)**	**.901 (.006)**	.921 (.015)	.916 (.009)	**.914 (.007)**	.823 (.006)	**.832 (.008)**	**.853 (.004)**	**.812 (.006)**	**.644 (.015)**	**.853 (.002)**

Table 5.3: Results (B-Cubed F-scores) with language families indicated across columns along with standard deviations in parentheses for cross-validated values. The best scores within a specific train-test split are shown in bold.

5.6 Results

The results are summarized in Table 5.3. The first column indicates the additional proportion of concepts that is moved from test data to training data. Thus, it roughly indicates the amount of increased supervision. The second column indicates the various methods discussed in §5.5.4 compared against the proposed model, *CogTran2*. The rest of the columns indicate the B-Cubed F scores (see §5.5.3) for various datasets discussed in §5.5.1. The last column indicates the mean B-Cubed F-scores averaged across the aforementioned datasets.

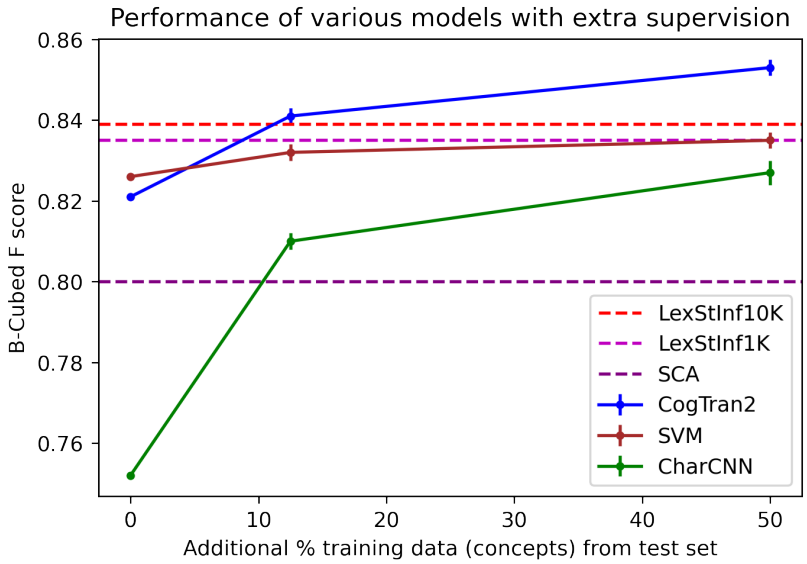


Figure 5.3: Plot of average B-Cubed F scores of various methods against % additional supervision

For the additional proportions +12.5% and +50%, the reported scores are means along with standard deviations (in parentheses) over the five validation sets (see §5.5.1). Note that the standard deviation for the overall averaged B-Cubed F score is considerably much less than those of individual datasets. This happens since in every run on a train-test split the model may perform high on one dataset or low on the other, yet when it comes to the mean performance it is quite stable. The averaged B-Cubed F scores are also plotted for each method across all the splits in Figure 5.3.

5.6.1 Discussion

From the results, it is visible that with increased supervision, CogTran2 improves consistently when compared to other supervised methods. At the same time, CogTran2 crosses the previous best LexStInf10K with additional +12.5% supervision, i.e., with only 20 concepts per family. Since the results of proportions +12.5% and +50% are cross-validated, it is possible to compare the performances throughout. Note that LexStat is not a supervised method and, hence, additional supervision does not make sense with it. With zero additional data, CogTran2 surpasses all the other methods on CHN and IE language families since they are present in training as well. While AN data is also present in both sets i.e., train and test, the individual languages do not overlap much as in the case of CHN

Method	Data Split		
	+0%	**+12.5%**	**+50%**
CogTran2	**0.821**	**0.841 (± 0.002)**	**0.853 (± 0.002)**
CogTran	0.815	0.830 (± 0.002)	0.841 (± 0.002)
CogTranL4	0.806	0.830 (± 0.002)	0.842 (± 0.004)
CogTranPair	0.779	0.813 (± 0.003)	0.833 (± 0.001)

Table 5.4: Mean B-Cubed F scores on various data splits for various ablation models. Standard deviations are indicated in parentheses for the data splits where cross-validation was performed.

and IE.

Although SVM beats CogTran2 on +0% additional data, which is not surprising since this is primarily dependent on LexStInf1K scores, it shows only a little increase in scores with an increase in additional training. Hence, overall, it is behind CogTran2 for the other two proportions. The maximum score of SVM does not appear to be significantly different from its base model LexStInf1K on whose scores it is dependent. We performed Student t-tests vis-à-vis SVM and CogTran2 scores for proportions +12.5% and +50%. On whatever dataset CogTran2 leads ahead of SVM, it is statistically significant for a 5% level of significance, i.e., $p < 0.05$. SVM leads ahead of CogTran2 significantly only on two datasets, namely, Austronesian (AN) and Romance (ROM) in both proportions. The reason for this is unclear as of now. Analysis with linguistic expertise in these languages could possibly unveil the cause.

CharCNN has the disadvantage of not using aligned input. Hence, it lags behind other models as expected (except SCA at extra supervision) despite showing a significant improvement over the additional training data.

Hence, it can be concluded that CogTran2 is the best performing model when there is sufficient labeled data. It is also likely to show improvement when there is plenty of labeled data. Further, given the availability of GPU and considering the present implementations, CogTran2 is much faster since it starts from MSA and not from independent pairwise computations.

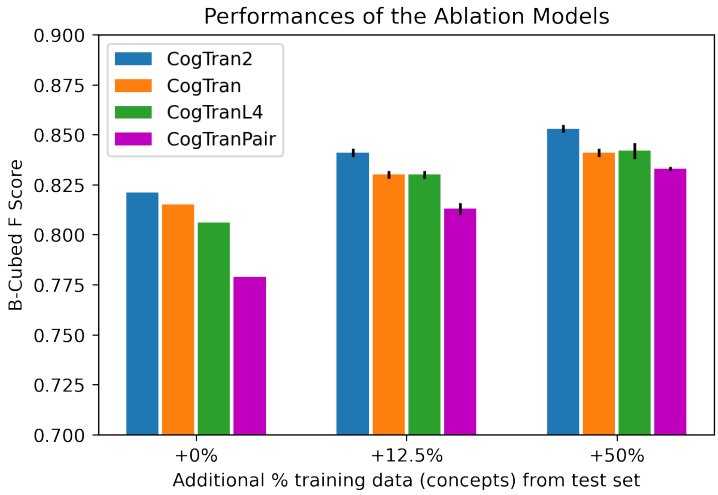


Figure 5.4: Mean B-Cubed F scores for various ablation models across the data splits.

5.6.2 Ablation Tests

The results of the ablation tests described in §5.5.4 on the data proportions +0%, +12.5% and +50% are presented in Table 5.4. The first column indicates the method and the second column lists the respective B-Cubed F-score averaged over all the datasets. These are mean scores along with standard deviations across all five cross-validated sets. These results are also plotted in Figure 5.4. CogTran, which lacks a Pairwise module (§5.4.4), underperforms significantly than CogTran2, which is the model proposed. Also, increasing the number of layers to 4 in CogTranL4 does not help either. Hence, it can be concluded that the Pairwise module alone contributes to further increasing the performance in CogTran2. Further since CogTrainPair, unlike the other two, starts from aligned word pairs akin to all other previous models, and takes input from an aligned word pair and outputs cognacy probability for that pair. Hence, the Pairwise module cannot be incorporated into this setup.

It is visible that CogTran, which acts on an MSA input performs way better than CogTranPair which acts on aligned word pairs. At the same time, CogTran (< 20 GPU min per split) is much faster than CogTranPair (about 1 GPU hr per split) for the same reason. In other words, let input MSA have r rows and c columns, then CogTranPair acts on all possible pairs of rows hence, in $O(r^2)$ steps. On the other hand, CogTran for a single MSA acts only once which results in the speed-up.

5.6.3 Error Analysis

To understand the working of CogTran2, we attempt to study some of the cluster predictions as follows. For this purpose, we consider CogTran2 trained on +12.5% proportion and the results on IE (Indo-European) dataset.

Sound Correspondences

The fundamental aspect for comparing two languages is to identify regular sound correspondences (Campbell, 2013). Methods like LexStat (List, 2012a) have built similarity metrics for cognacy judgement between two words giving weightage to both the recurrent sound correspondences as well as phonetic information. In this regard, we note that CogTran2 appears to have learned some recurrent sound correspondences by observing the initial consonant. For example, Proto-Indo-European **s-* undergoes lenition in Hellenic branch and appears as *h-* is Ancient Greek (Mallory and Adams, 2006). In the dataset we have used, two words occur as instances for this sound change, namely, /ˈhɛːlios/ 'sun' and /ˈhals/ 'salt'. Both these words are clustered correctly with their cognates in other daughter languages such as Old Norse /soːl/, Oriya /surdʒɔ/ in case of the concept 'sun' and English /sɔːlt/, French /sɛl/ in case of the concept 'salt'. Thus, one may assume that the sound change PIE **s* > Ancient Greek *h* has been learned by the model.

Another set of sound changes where position of articulation changes is Grimm's law where Proto-Indo-European hard consonants undergo a chain shift in Germanic family (Mallory and Adams, 2006). For instance, in the velar shift defined by Grimm's law i.e., $*g^h > *g > *k > *h$, change in the place of articulation occurs in the sound change **k* > **h*. The model also learns this sound change as supported by the instances mentioned as follows. For the concept 'dog', German /hʊnt/ has been correctly clustered together with Ancient Greek /ˈkyɔn/ and Old Irish /kuː/. Further, for the concept 'horn', German /hɔrn/ and Ancient Greek /keras/ are similarly clustered together correctly. This sound change has been learned by the model to an extent that unrelated German /hɪmḷ/ and Latin /kae̯lʊm/ meaning 'sky' have been classified as cognates. Both the sound changes mentioned above have two instances as examples in the dataset.

On the other hand, Marathi /dzaŋ/ and Ossetic /zon/ for the concept ‘know’ have been incorrectly classified as different. This happens to be the only example where the phonemes /dz/ and /z/, which fall in different sound classes, co-occur in the respective languages. Hence, it may be concluded that at least two examples are needed to learn a sound change. However, it is desirable to perform a thorough quantitative analysis of recurrent sound changes to support these findings. It could not be performed due to a lack of readily available annotated data for the same.

Partial Cognacy

Further, the network seems to consider the entire word and not just the important root in some cases. For example, for the meaning ‘woman’, Old Norse /kvenː maðr/ and Icelandic /kʰvɛn maːðr/ have been assigned a different cluster than that of Old Swedish /kvinːa/ and Danish /gʰvenə/. This is conceivable since affixes cannot be learned to be ignored easily. Detection of sub-word cognates in presence of such affixes is part of *partial cognacy* problem which was dealt in List et al. (2016). It is, thus, clear that CogTran2, at its present training level, cannot distinguish partial cognates.

Other Errors

Many errors are, however, somewhat incomprehensible. For example, in the case of ‘tooth’, Greek /ˈðondi/ has been clustered together with English /tuːθ/ but not with Italian /dɛntɛ/. There could be a role of root vowel in this particular example. Nevertheless, it is important to understand the source of errors which demands linguistic expertise to identify the bottlenecks of the current models and to improve beyond them.

5.7 Summary

In this chapter, we have proposed a Transformer-based model inspired by evolutionary biology for the task of automatic cognate detection. The model can harness efficiently the labeled data and consequently, with sufficient data, outperforms existing approaches that do not make efficient use of supervision data. In particular, better results are obtained

with only 20 concepts per family on some of the datasets. To the best of our knowledge, we proposed for the first time in this particular problem a method that directly outputs link probabilities, i.e., pairwise similarities from an input MSA in an end-to-end fashion, unlike all the previous methods which act on aligned pairs of words. We demonstrated through the primary results and ablation studies that this approach of inputting MSA rather than paired alignments results not just in a significant increase in performance but also in drastically reducing the computation time. We have also demonstrated by observing few outputs that the model is capable of learning regular sound changes from just two example instances in the data for a particular sound change.

Evaluation of Cognate Transformer on phylogenetic reconstruction task (Rama et al., 2018) is an unexplored problem and, thus, can be a potential topic of future work.

Limitations

As mentioned in §5.6, the proposed model lags on the datasets Romance and Austronesian somewhat behind SVM and LexStat-Infomap and on Pama-Nyungan concerning Lexstat-Infomap despite increasing the supervision. While the performance on the Romance dataset is near saturated (>92%) in any case, the lag in performance on Austronesian and Pama-Nyungan data is an issue that is required to be studied with domain linguistic expertise to understand the bottleneck of this model. Similarly, although our model improves drastically on Sino-Tibetan by 5% when compared to the previous best, it is an underperforming dataset since the B-Cubed F-scores on all other datasets except this are more than 80%. Thus, a similar study with linguistic expertise is required to identify the bottleneck of the overall methodologies. Additionally, as mentioned in §5.5.2, a GPU memory of 10GB could only accommodate a batch of size 4 during training with maximum MSAs, i.e., when the number of languages in a family was 136. Thus, larger GPU storage is required for larger mass comparisons involving more languages under comparisons. As mentioned in §5.6.3, the ability of the model to learn regular sound correspondences has only been determined by anecdotal instances. A more thorough quantitative study is desirable, which requires annotated data for the same. The model also does not account for partial cognacy,

i.e., identifying distinctions between exact cognates versus morphologically modified or compounded cognates (see §5.6.3) as addressed in List et al. (2016). Further, the model is also not tuned at this point to distinguish between true cognates and borrowals.

Chapter 6

Likelihood Ratio Test of Genetic Relationship

Lexical resemblances among a group of languages indicate that the languages could be genetically related, i.e., they could have descended from a common ancestral language. However, such resemblances can arise by chance and, hence, need not always imply an underlying genetic relationship. Many tests of significance based on permutation of wordlists and word similarity measures appeared in the past to determine the statistical significance of such relationships. We demonstrate that although existing tests may work well for bilateral comparisons, i.e., on pairs of languages, they are either infeasible by design or are prone to yield false positives when applied to groups of languages or language families. To this end, inspired by molecular phylogenetics, we propose a likelihood ratio test to determine if given languages are related based on the proportion of invariant character sites in the aligned wordlists applied during tree inference. Further, we evaluate some language families and show that the proposed test solves the problem of false positives. Finally, we demonstrate that the test supports the existence of macro language families such as Nostratic and Macro-Mayan. This chapter is based on Akavarapu and Bhattacharya (2024b).

6.1 Introduction

Languages that descend from a common ancestral language are termed to be *genetically related*. The existence of lexical resemblances between the two languages is a preliminary indication that they could be related. Such resembling lexicons that truly have a common origin are called *cognates*. For instance, Sanskrit *nāma* and English *name* are cognates that can be traced to Proto-Indo-European *$*h_3$nómn*. However, such resemblances can also occur out of sheer chance. For instance, Persian *bad* and *behtar* accidentally resemble English *bad* and *better* respectively, but are not true cognates[1]. Hence, it is necessary to show *statistical significance* on any appropriate measure that captures the lexical relatedness before arguing for a genetic relationship among any group of languages or language families (Campbell, 2013).

Several significance tests appeared in the past to address this problem, with the majority of them based on permutation tests, starting from Oswalt (1970). Given wordlists of a group of languages to be evaluated for a genetic relationship, these tests obtain the null distribution of a certain measure capturing similarity between word pairs by random permutations of the wordlists. Such tests either act *bilaterally*, i.e., on a pair of languages or proto-languages, or *multilaterally* on a group of languages. Among these, the multilateral comparison, which was made famous by Greenberg (1963, 1971, 1987, 2000) in traditional historical linguistics, has been a subject of much criticism (Poser and Campbell, 2008). Hence, the preferred way of comparing two language families has been to compare their reconstructed proto-forms bilaterally. However, Greenberg (2005) argues that genetic classification should precede proto-language reconstruction. Moreover, there is often a lack of agreement on reconstructed proto-forms both in terms of phonology and semantics which gives room for sufficient manipulation of wordlists that can in turn alter the results of significance tests (Kessler, 2015). Further, we demonstrate that multilateral permutation tests (Kessler and Lehtonen, 2006; Kessler, 2007) yield false negatives even after incorporating complex word similarity metrics such as SCA and LexStat (List, 2010, 2012a).

[1]Persian *bad* is of uncertain origin while *behtar* ultimately derives from PIE *$*h_1$wésus*. On the other hand, English *better* derives from PIE *$*b^h$edrós* and is cognate with Sanskrit *bhadrá*

To overcome these issues, we turn to *phylogenetic analysis* (Wiley and Lieberman, 2011) that is known to approximately capture the ancestral states and has been applied to phonological reconstruction tasks such as proto-language and cognate reflex prediction tasks (Jäger, 2019, 2022) with reasonably good results. Specifically, we propose a *likelihood ratio test* (LRT) where we expect the difference in likelihoods of the best trees under null and alternate hypotheses to capture genetic relatedness. The null hypothesis assumes negligible proportion of invariant sites while the alternate hypothesis assumes significant proportion of invariant sites. Intuitively, related languages should have more positions where a character or a sound class is invariant than unrelated languages. Hence, we essentially capture the notion of relatedness as possessing a relatively high proportion of invariant sites. Further in this test, reconstructed proto-forms are not required and at the same time, the evolutionary tree structure is strictly imposed by design, unlike the multilateral model, thereby effectively circumventing the aforementioned methodological problems. Although inspired by similar tests from molecular phylogenetics, the test we propose is novel in the sense that the problem of testing common descent never arises in biology since monogenesis is accepted as a fact therein (Kessler, 2008). We further evaluate the test on various language families and demonstrate that the test does not misclassify unrelated languages as related.

We finally show that the test supports the existence of the macro-families Nostratic (Bomhard and Kerns, 1994) and Macro-Mayan (Campbell, 1997). While such an attempt to justify the existence of macro-families using bootstrap analysis of distance-based phylogeny is found in Jäger (2015), expressing statistical significance in terms of likelihood ratio is preferred over bootstrap support values whose interpretation is debated in molecular phylogenetics (Anisimova and Gascuel, 2006).

Our contributions are summarized as follows.

- We have proposed a *likelihood ratio test* to determine the *genetic relatedness* of a group of languages based on *invariant site proportions*.
- We have demonstrated by applying various language sets that the test does not exhibit the problem of false positives nor requires reconstructed proto-forms, unlike the

previously proposed tests.

- We have found through the test some supporting evidence for the existence of macro-families namely Nostratic and Macro-Mayan

The rest of the chapter is summarized as follows. Related work is discussed in §6.2. The methodology of the test is presented in §6.3. Evaluation details such as datasets and details of previous methods and variants are discussed in §6.4. The results are discussed in §6.5. The application of the method on long-range comparisons is discussed in §6.6. This chapter is finally concluded in §6.7.

6.2 Related Work

Permutation test for bilateral language relationship comparisons was introduced by Oswalt (1970). The significance of sound correspondences by brute force probability calculation was proposed by Ringe (1992, 1996). This approach was however criticized for not being able to show significance for known related pairs of languages like Latin-English and also for accounting phonologically implausible sound correspondences (Kessler, 2001). Multilateral permutation tests were proposed by (Kessler and Lehtonen, 2006; Kessler, 2007). Several applications of permutation tests exist such as (Turchin et al., 2010; Kassian et al., 2015).

Some notable likelihood ratio tests in molecular phylogenetics, mostly on topologies, include (Huelsenbeck and Bull, 1996; Huelsenbeck et al., 1996; Goldman et al., 2000; Anisimova and Gascuel, 2006) where bootstrap analysis is argued to be not so optimal to establish statistical significance on phylogenies. Otherwise, support for macro-families through bootstrap analysis for distance-based trees is shown in Jäger (2015). Comparisons of various methods of phylogenetic reconstruction such as distance-based and binary-character-based are given by Jäger (2018). Sound-class character-based phylogenetic analysis is found in (Jäger, 2019, 2022). Usually, Bayesian phylogenetic inference on binary cognate encodings gives good results (Rama et al., 2018; Rama and List, 2019).

Although the likelihood ratio metric is common for both past and present-day lan-

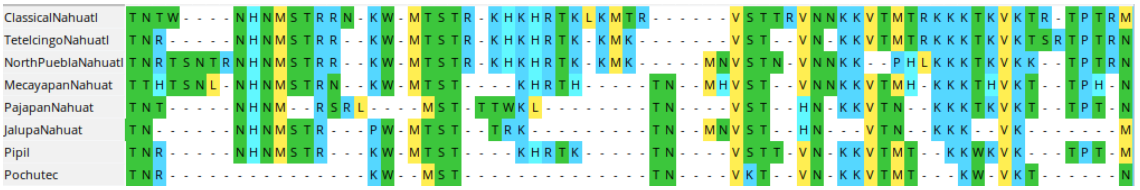


Figure 6.1: A section of character matrix for Uto-Aztecan family consisting of concatenated Multiple Sequence Alignments (MSAs) of consonant classes, one from each concept

guage models, the utility of this test using invariant sites outside computational historical linguistics is unknown.

6.3 Methodology

The key concept revolves around the idea that any hypothesis, in this case, a hypothesis on a phylogeny, is preferred over a competing null hypothesis if it is significantly more likely, i.e., has a higher likelihood than the latter. Given the wordlist data encoded as an aligned character matrix, related languages are expected to have a higher number of *invariant* columns. Thus, our null hypothesis consists of a phylogeny with a small proportion (fixed at 1%) of invariant sites, whereas the alternative hypothesis consists of a phylogeny with a larger but reasonable proportion (fixed at 6%) of invariant sites. The observed difference in their likelihood of real data is compared with that of data simulated from the null hypothesis through parametric bootstrapping and, accordingly, one of the hypotheses is rejected. The steps are elaborated next.

6.3.1 Character Matrix

The wordlists of a given group of languages, as mentioned previously, are encoded in the form of a *character matrix*. It consists of concatenated aligned words per concept, i.e., meaning. Thus, each row represents a language or *taxon*, and each column, also referred to as *site* in this chapter, consists of phoneme classes, e.g., Dolgopolsky classes. Use of phoneme classes facilitate comparisons across languages which may otherwise be difficult due to the existence of significant variation in the IPA consonants. Formally, let the input language set be $\{L_1, \ldots, L_m\}$, whose genetic relatedness is to be verified statistically. Let

Greek_Anc	K	R	-	S	
Latin	K	R	N	-	-
English	H	R	N	-	-
Sanskrit	S	R	N	K	-

Table 6.1: Example of a Multiple Sequence Alignment (MSA) of consonant classes for a single concept 'horn'.

there be n concepts $C_1, \ldots, C_n$ in the wordlists. Each language L_i should have for each concept C_j a single word, say w_{ij}. If a language has multiple words for a single semantic slot, only the one with fundamental or core meaning is observed manually and retained, following the recipe by Kessler (2001). For instance, if the meaning 'dull' has words *dull* and *unsharp*, *dull* is of core or fundamental meaning. Another example would be for the meaning 'belly', Latin *venter* is more fundamental than *abdōmen*. If it so happens that it still remains unresolved after this step (very few cases), a single word is randomly picked up. In case a language has no word for a semantic slot, it is represented as a gap '–'. For each concept C_j and alphabet set $\mathbb{A}$, let $W^j \in \mathbb{A}^{m \times l_j}$ represent a multiple sequence alignment (MSA) of words where l_j is the length or the number of phonemes with vowels removed[2] in each word. The final character matrix $X \in \mathbb{A}^{m \times N}$ is concatenation of W^j, i.e., $[W^1 \ldots W^n]$ across columns and $N = \sum_{j=1}^{n} l_j$.

For example, consider a cognate set meaning 'horn' from a few Indo-European languages namely, Ancient Greek *keras*, Latin *cornu*, English *horn*, and Sanskrit *śṛṅga*. The resultant character matrix for this single meaning is a multiple sequence alignment with vowels removed and consonants encoded as Dolgopolsky classes as illustrated in Table 6.1. The final character matrix is the concatenation of such matrices across all the concepts. For an illustration of a final character matrix, see Figure 6.1, which is generated by MEGA11 (Tamura et al., 2021). In general, multiple sequence alignment is a fundamental step in several state-of-the-art methods in computational historical linguistics (Akavarapu and Bhattacharya, 2023a, 2024a).

[2]Since the root form CVC is universal, including vowels results in spurious relationships. Further, languages of Caucasus like Georgian are rich in consonant clusters and, as a result, comparing them to others becomes difficult when vowels are considered.

6.3.2 Substitution Model

A *substitution model* describes the evolution of a character at a site assuming a Markovian process. Various substitution models have been described for various alphabets such as nucleotides, amino acids, etc. In this chapter, we assume the simplest possible model where substitution rates are assumed to be equal between all the pairs of distinct characters. The resultant model is known as the Jukes-Cantor model (Jukes et al., 1969) in case of nucleotide substitutions and as Poisson (Bishop and Friday, 1987) in case of amino-acid substitutions. Formally, let the number of characters in the alphabet $\mathbb{A}$ be N. An element q_{ij} of the rate matrix Q, which denotes the rate at which character i mutates to character j is defined as follows:

$$q_{ij} = \mu \cdot \pi_i \text{ , } i \neq j \text{ (equal rates)} \tag{6.1}$$

where π_i denotes the frequency of character i at the site and μ is the rate of mutation. The diagonal element should satisfy the normalization constraint:

$$q_{ii} = -\sum_{j \neq i} q_{ij} \tag{6.2}$$

The probability of transition $i \rightarrow j$ in time t is given by the matrix $P(t) = \{p_{ij}\} = e^{Qt}$. Likelihood of an evolutionary tree with topology T can be, thus, calculated from the substitution matrix where branch lengths V would denote the time. See Figure 6.2 for an example.

6.3.3 Maximum Likelihood Tree (ML-tree)

For any phylogenetic tree with topology T, branch lengths V, other parameters such as shape parameter of heterogeneous rate, the proportion of invariant sites denoted by Θ, and with the observed data i.e., character matrix X, the *likelihood* is defined as the product of likelihoods at each site as given by the following equation, assuming independence for

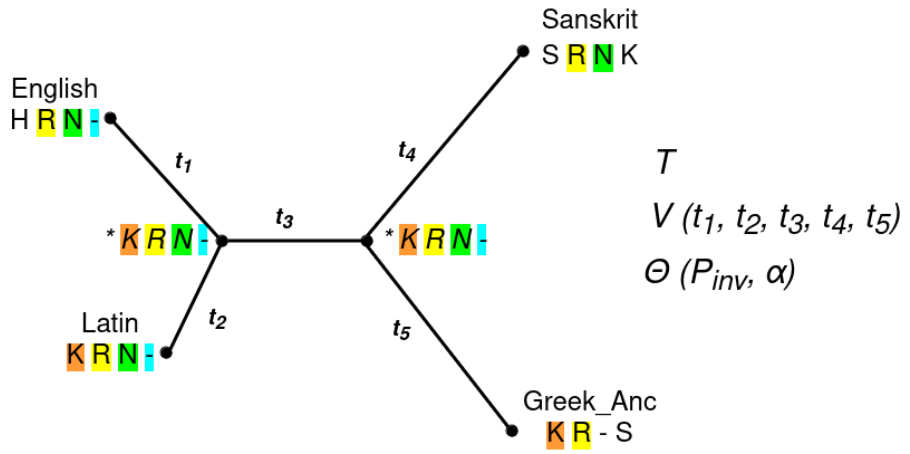


Figure 6.2: Likelihood of a phylogenetic tree T with branch lengths V is computed assuming equal rate evolution substitution model.

simplicity:

$$\mathcal{L}(T, V, \Theta|X) = \prod_{i=1}^{N} P(X_i|T, V, \Theta) \tag{6.3}$$

The site independence assumption also restricts the number of parameters. Given the limited amount of data, which is restricted to 100-200 wordlists[3], this is, thus, more suitable. Complex models such as bigram-based ones may be employed if sufficient data is available.

The parameters that maximize the likelihood, $\hat{T}, \hat{V}$, and $\hat{\Theta}$, define the *maximum likelihood tree* which is usually obtained by heuristic search in the parameter space. Typically, a tree is initialized either randomly or by some heuristic means, and from there, the tree space is explored through tree modifying operations to get the "best" tree. For a given tree, likelihood is computed using the well-known Felsenstein's pruning algorithm from phylogenetics (Felsenstein, 1973, 1981).

6.3.4 Invariant Sites

Invariant sites are those sites that are constant or evolve very slowly. These can be estimated through a maximum likelihood search along with other parameters. The proportion of invariant sites, P_{inv} may be known beforehand or estimated. Given the invariant sites, the

[3]For any language, 100-200 basic wordlists tend to have least presence of borrowings.

likelihood defined in §6.3.3 is only the product of likelihoods across the variant sites.

Our observation is that estimated P_{inv} is higher (>0.06) among related languages while lower (≈0.01) among (possibly) unrelated languages. Based on this observation and preliminaries, we now describe the likelihood ratio test.

6.3.5 Likelihood Ratio Test (LRT)

Given a null hypothesis H_0 and a competing alternative hypothesis H_a, the latter is preferred if it is more likely than the former i.e., $\mathcal{L}_{H_a} > \mathcal{L}_{H_0}$. In our case, the hypotheses consist of respective phylogenetic tree parameters estimated for ML-trees, i.e., H_0 consists of $\hat{T}_0, \hat{V}_0, \hat{\Theta}_0$ and H_a consists of $\hat{T}_a, \hat{V}_a, \hat{\Theta}_a$. The likelihood ratio test defines the following metric to decide whether to reject the null hypothesis:

$$\delta = 2 \cdot \ln \left(\frac{\mathcal{L}(\hat{T}_a, \hat{V}_a, \hat{\Theta}_a)}{\mathcal{L}(\hat{T}_0, \hat{V}_0, \hat{\Theta}_0)} \right) \tag{6.4}$$

The *Likelihood Ratio Test* (LRT) metric δ was shown to asymptotically follow a chi-squared distribution when the null hypothesis is assumed with the degrees of freedom $p-q$, where p and q respectively are the numbers of free parameters in the alternate and the null hypotheses (Wilks, 1938). However, it was argued that this may not hold in general for phylogenetic problems due to the discrete nature of tree topology (see (Huelsenbeck and Bull, 1996; Huelsenbeck et al., 1996; Anisimova and Gascuel, 2006) for relevant work). As a result, the distribution of δ is determined by a parametric bootstrapping method where it is measured on the data simulated by the parameters estimated assuming the null hypothesis H_0 to hold, i.e, using the parameters $\hat{T}_0$, $\hat{V}_0$ and $\hat{\Theta}_0$.

As mentioned in §6.3.4, we propose LRT to test the relatedness of a group of languages using varying proportions of invariant sites. In other words words the null hypothesis H_0 consists of invariant site proportion P^0_{inv} and alternate hypothesis H_a consists of P^a_{inv} where $P^0_{inv} < P^a_{inv}$ as per the observations discussed in §6.3.4.

The typical way of obtaining the distribution for δ under H_0 involves finding the parameters $\{\hat{T}_0, \hat{V}_0, \hat{\Theta}_0\}$ and $\{\hat{T}_a, \hat{V}_a, \hat{\Theta}_a\}$ for the best trees respectively under H_0 and H_a

Family	Abbrv.	Languages	Concepts	Words
Afrasian	AfA	21	39	770
Dravidian	Drav	4	183	716
Indo-European	IE	12	185	2209
Kartvelian	Kart	1	180	180
Lolo-Burmese	LoBur	15	39	565
Mayan	May	30	94	2667
Mixe-Zoque	MZ	10	94	905
Mon-Khmer	MKh	9	199	1701
Mon-Khmer	MKh	16	94	1332
Munda	Mun	4	199	759
Uto-Aztecan	UAz	9	94	803

Table 6.2: Language families considered in this study.

along with observed δ, say $\hat{\delta}$. Further, several, say k, bootstrap replicates are generated from the topology, branch lengths, and other parameters defined by $\{\hat{T}_0, \hat{V}_0, \hat{\Theta}_0\}$, i.e., assuming H_0. Next, the maximum likelihood search is run again on these replicates to obtain several samples for δ, say $\{\delta_1, \ldots, \delta_k\}$. However, we found considerable variation in $\hat{\delta}$, since the maximum likelihood search is only a heuristic and is affected by initialization. As a result, we obtain several samples for $\hat{\delta}$, say $\{\hat{\delta}_1, \ldots, \hat{\delta}_k\}$ by running the search k times and based on the null parameters, a single bootstrap replicate is generated for each search to consequently obtain $\{\delta_1, \ldots, \delta_k\}$ for corresponding k searches. Finally the *p-value* for $\mathbb{E}[\delta] < \mathbb{E}[\hat{\delta}]$ is obtained by one-sided paired t-test. If the p-value is less than a threshold (usually 0.05), we conclude that H_a may hold or, in other words, there are at least P^a_{inv} proportions of sites that are significantly invariant and, thus, the languages under consideration are likely to be related.

6.4 Experimental Setup

The section discusses the details of the experiments including datasets, baseline models, and implementation details.

6.4.1 Datasets

The data for evaluating the tests consists of wordlists from multiple language (sub-)families and their combinations. Combinations of related sub-families serve as positive examples while those of unrelated serve as negative examples. Evaluating the macro-families also consists of language groups whose relationship is only distantly suggested such as Nostratic (Bomhard and Kerns, 1994).

The details of data from each family are shown in Table 6.2. Out of these, Mon-Khmer and Munda (200 wordlists) are extracted from the Austro-Asiatic data from Rama et al. (2018). Data for Old languages of Nostratic comprising Indo-European, Dravidian, and Kartvelian are prepared by us from the Swadesh 200-wordlists available at Wiktionary[4]. Data for all the other families are obtained from Rama (2018) which were, in turn, collected from various publicly available sources. The datasets are the same as those found in related tasks such as automated cognate detection and proto-language reconstruction.

In the Nostratic grouping, we considered the languages that are surviving or have surviving descendants and were attested by the 10th century CE. The motivation behind this choice is that older languages should be closer to the ancestral language and each other if at all there is any relationship. Several languages including literary Dravidian languages, Georgian, and Armenian are mostly conservative and deviate little from their old forms. The data is pre-processed by excluding motivated word forms including onomatopoeia, and nursery forms, listed in Kessler (2001). Short forms, i.e., words consisting of single syllables are also excluded. Such cleaning is necessary to avoid the appearance of spurious relationships. In the case of Nostratic, we were also careful to exclude borrowings by tracing etymologies from Wiktionary[4]. This step could not be extended to other language families due to a lack of readily available etymological information.

All the methods employed in this work, including both the proposed one and baseline ones described in §6.4.2, involve the construction of a phylogenetic tree. Hence, we also compare the methods on a tree construction task where we see how well the trees match the golden truth trees wherever available. The data for this task is taken from Rama et al.

[4]https://en.wiktionary.org/wiki/Category:Swadesh_lists_by_language

Family	Abbrv.	Languages	Concepts	Words
Austro-Asiatic	AA	58	200	11001
Austronesian	AN	45	210	8309
Indo-European	IE	42	208	8478
Pama-Nyungan	PN	67	183	11503
Sino-Tibetan	ST	64	110	6762

Table 6.3: Language family datasets for tree construction.

(2018) as summarized in Table 6.3.

6.4.2 Multilateral Permutation Test

As mentioned in §6.1, most previous methods compare languages bilaterally, i.e., a pair at a time. As a result, the only possible way to compare the language families in this approach is to compare their reconstructed proto-languages. However, proto-forms of a proto-language are not often universally agreed which leads to considerable allowance of manipulation that can affect the results (Kessler, 2015). An alternate solution to determine the significance of the relationship among multiple languages was proposed by Kessler and Lehtonen (2006) and Kessler (2007) who employ a permutation test based on multilateral comparison. This has been well received in historical linguistics (Ringe and Eska, 2013).

The test is based on nearest-neighbour hierarchical clustering where at any point two closest clusters are lumped into one cluster. The basic distance measure, $\hat{d}(A, B)$, between any two clusters A and B is the average of distances between all possible pairs of languages in these clusters, i.e.,

$$\hat{d}(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \tag{6.5}$$

where the distance $d(a, b)$ between any two languages a and b is the mean distance between the pairs of words over all concepts. Following the notations of §6.3.1 where w_{aj} and w_{bj} are words in languages a and b respectively from concept C_j,

$$d(a, b) = \frac{\sum_{C_j, w_{aj} \neq \emptyset, w_{bj} \neq \emptyset} d(w_{aj}, w_{bj})}{|\{C_j : w_{aj} \neq \emptyset, w_{bj} \neq \emptyset\}|} \tag{6.6}$$

Taking an average over all languages essentially enforces multilateral comparison, i.e., multiple languages are being considered equally to compute the outcome. Further, the algorithm thus described is the same as UPGMA tree construction method (Sokal and Michener, 1958) where at any bifurcating node, a uniform rate of evolution is assumed across daughter clades. The final similarity metric $\hat{s}(A, B)$ is determined by the following statistic that is computed based on a random permutation of words across each column (taxon) which yields random distances $d(A, B)$:

$$\hat{s}(A, B) = \frac{\mathbb{E}[d(A, B)] - \hat{d}(A, B)}{\mathbb{E}[d(A, B)]} \tag{6.7}$$

The *p-value* of two language clusters A and B is the frequency of the event $\hat{d}(A, B) \geq d(A, B)$ relative to the total number of random permutations. Language clusters A and B are considered to be *related* if the p-value is less than 0.05. The given languages are termed *related* if the final two clusters that are merged at the root are related (Kessler and Lehtonen, 2006).

Kessler (2007) ran this test using various word similarity metrics which almost give similar results. Among these metrics, we ran on P1-dolgo which is a binary metric that determines whether the consonant class of the word's initial consonant matches or not. Additionally, we employ the binary similarity measure introduced by Turchin et al. (2010) to test the significance of the Altaic family where the first two consonants are considered. We further test continuous word distances introduced by List (2010) (SCA) and List (2012a) (LexStat) that are based on sequence alignment techniques which were introduced in the context of automated cognate detection.

6.4.3 Implementation

We mapped the consonant classes to the protein alphabet since phylogenetic software expects input as either nucleotide or amino acid sequences. Moreover, most of the amino acid letters and Dolgopolsky classes are identical. In this regard, there is only one exception, namely, 'J' which is absent in the former but present in the latter and is, hence, simply

Method	MKh	Mun	MKh-Mun	IE	Drav	May	MZ	UAz	MKh-May	MKh-UAz	AfA-LoBur
Related	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
P1-Dolgo	0.123 (<0.001)	0.243 (<0.001)	0.080 (<0.001)	0.071 (<0.001)	0.440 (<0.001)	0.228 (<0.001)	0.412 (<0.001)	0.572 (<0.001)	0.007 (<0.001)	0.005 (0.063)	0.017 (<0.001)
Turchin	0.019 (<0.001)	0.124 (<0.001)	0.019 (<0.001)	0.028 (<0.001)	0.292 (<0.001)	0.126 (<0.001)	0.256 (<0.001)	0.402 (<0.001)	0.003 (<0.001)	0.003 (0.005)	0.004 (<0.001)
LexStat	0.065 (<0.01)	0.138 (<0.01)	0.048 (<0.01)	0.036 (<0.01)	0.197 (<0.01)	0.129 (<0.01)	0.244 (<0.01)	0.306 (<0.01)	0.028 (<0.01)	0.018 (<0.01)	0.033 (<0.01)
SCA	0.087 (<0.01)	0.187 (<0.01)	0.074 (<0.01)	0.056 (<0.01)	0.296 (<0.01)	0.177 (<0.01)	0.304 (<0.01)	0.400 (<0.01)	0.015 (<0.01)	0.006 (<0.01)	0.031 (<0.01)
LRT	9.205 (<0.001)	1.58 (<0.001)	14.18 (<0.001)	26.154 (<0.001)	1.78 (<0.001)	68.212 (<0.001)	7.192 (<0.001)	10.448 (<0.001)	-14.359 (0.280)	-12.188 (0.065)	-10.768 (0.979)

Table 6.4: Significance testing on various existent and non-existent families. The values indicate the similarity measure $\hat{s}$ in the case of permutation tests and in the case of LRT they indicate the mean of statistic $\hat{\delta}$. Values in parentheses indicate p-value. False positives are marked in red.

replaced with 'I', which is in turn absent in Dolgopolsky classes. The multiple sequence alignments are obtained from CLUSTALW2 (Larkin et al., 2007) while the best trees and their corresponding likelihoods were computed using IQ-TREE (Nguyen et al., 2015). As described in §6.3.4 and §6.3.5, the proportions of invariant sites P^0_{inv} and P^a_{inv} are set to 0.01 and 0.06 respectively for null (H_0) and alternate (H_a) hypotheses. The parametric bootstrap replicates are generated using AliSim (Ly-Trong et al., 2022), an extension of IQ-TREE. To replicate as closely as possible, gaps present in the original character matrices are retained in the replicates. We calculate the p-value based on a sample size of $k = 15$. The outcomes are observed to be stable beyond this size. The word similarity metrics used in the baseline models are computed by using Lingpy (List and Forkel, 2021). For the phylogenetic tree construction task, MEGA11 (Tamura et al., 2021) was used to deduce the maximum likelihood tree (ML-tree) with the aforementioned model with an additional gamma rate heterogeneity parameter with two distinct rates whose shape is estimated. We name this method *ML-P+I+G2*.

The *generalized quartet distances* (GQD) (Pompei et al., 2011) between the predicted and the gold trees are computed from quartet distances obtained using qdist (Mailund and Pedersen, 2004). The *quartet distance* between two trees measures the number of four-leaf-subsets that have dissimilar topologies. Unlike biological phylogenetic trees, language trees are often multifurcated. Hence, GQD excludes penalties over the order of bifurcations. The code and relevant data have been made publicly available[5]. Further implementation details can be found in `README.md` therein.

[5]https://github.com/mahesh-ak/PhyloVal

6.5 Results

The primary results are tabulated in Table 6.4, where the results of LRT (last row) are compared with those of the multilateral permutation tests. Except for LRT, the column 'Method' indicates the distance metric employed in the permutation test. The row 'Related' indicates the current consensus about the relatedness of the language families. For the permutation test, the values indicate the similarity metric $\hat{s}$ defined in Eq. (6.7), as measured at the root. On the other hand, for LRT the values indicate the mean of observed $\hat{\delta}$ (see §6.3.5). The p-values are indicated in parentheses. The standard threshold of 0.05 is assumed for p-values. Please refer to Table 6.2 and Table 6.3 for abbreviations of various language families.

One can observe that false positives, indicated in red, are absent for LRT, in contrast with multilateral permutation tests which exhibit false positives in all cases (except P1-Dolgo for MKh-UAz). However, we note that the similarity scores of the Turchin measure are consistently small (< 0.005) for negatives irrespective of the significance implied by the p-value. Hence, it may be noted that Turchin could be a good measure for permutation tests when similarity scores are taken into consideration.

Further, one can observe from Table 6.4 that mean $\hat{\delta}$ values are small for valid families such as Mun and Drav. This has to do with the fact that the data for these families consists of a lower number of taxa (see Table 6.2). Hence, althought the $\hat{\delta}$ measure need not imply strength, its sign implies which hypothesis is to be preferred, i.e., the one with a larger proportion of invariant sites in case of a positive value and the one with a smaller proportion of invariant sites in case of a negative value.

6.5.1 Tree Construction

As mentioned in §6.4.1, both the methods output a tree, and, therefore, the methods have been evaluated on the tree construction task. The purpose of this task is to ensure that the proposed methods have indeed a good sense of phylogenetic inference and are, hence, appropriate to carry out significance tests over phylogenies. The results are tabulated

Method	AA	AN	IE	PN	ST	Avg
P1-Dolgo	0.060	0.208	0.033	0.175	0.188	0.133
Turchin	0.069	0.195	0.058	0.175	0.275	0.154
LexStat	0.051	0.178	**0.020**	0.164	0.096	0.102
SCA	0.049	0.119	0.025	0.166	**0.087**	0.089
ML-P+I+G2	**0.026**	**0.065**	0.033	**0.145**	0.125	**0.079**

Table 6.5: Comparison of the methods on phylogenetic tree construction task provided as GQD scores. The best results are in **bold**.

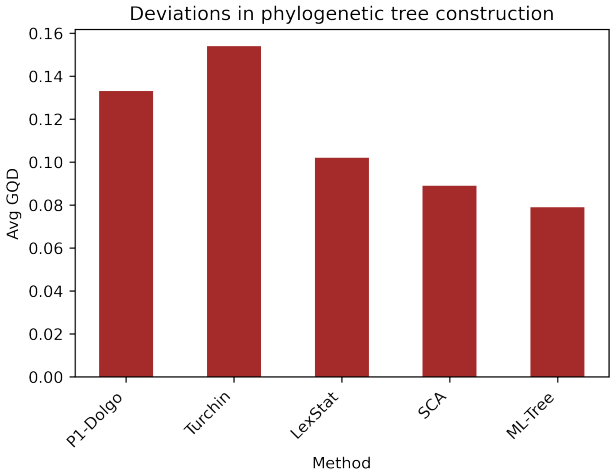


Figure 6.3: Mean GQD plotted for various methods

in Table 6.5. The mean GQD scores are plotted in Figure 6.3. By comparing with the mean scores of state-of-the-art language phylogeny inference methods on this data, ML-P+I+G2 (0.079) is a few steps behind Bayesian inferred tree (0.066) (Rama et al., 2018) and maximum a posteriori tree (0.051) (Rama and List, 2019). Hence, it can be concluded that consonant-class-based character matrix encoding is almost as good as cognate-based binary character matrix encoding while probabilistic methods based on character matrices are superior to distance-based methods for this task. Among the distance-based approaches, one with the SCA metric performs best. A similar situation was observed in Rama et al. (2018) and Rama and List (2019) where SCA-based cognates yield the best performance. However, it should be noted that SCA and LexStat-based measures yield false positives on significance testing (Table 6.4) despite their performance on this task. The low GQD scores across the methods in general demonstrate that the 100-200 wordlists are indeed good enough to infer language phylogenies.

Method	Drav-IE	Drav-IE-Kart	May-MZ	May-UAz	May-MZ-UAz
P1-Dolgo	0.046 (<0.001)	0.038 (<0.001)	0.033 (<0.001)	0.046 (<0.001)	0.036 (<0.001)
Turchin	0.017 (<0.001)	0.002 (0.197)	0.012 (<0.001)	0.012 (<0.001)	0.008 (<0.001)
LexStat	0.024 (<0.01)	0.014 (<0.01)	0.033 (<0.01)	0.027 (<0.01)	0.024 (<0.01)
SCA	0.024 (<0.01)	0.007 (0.01)	0.019 (<0.01)	0.024 (<0.01)	0.015 (<0.01)
LRT	24.882 (<0.001)	0.316 (<0.001)	20.988 (<0.001)	-1.035 (<0.001)	-9.819 (<0.001)

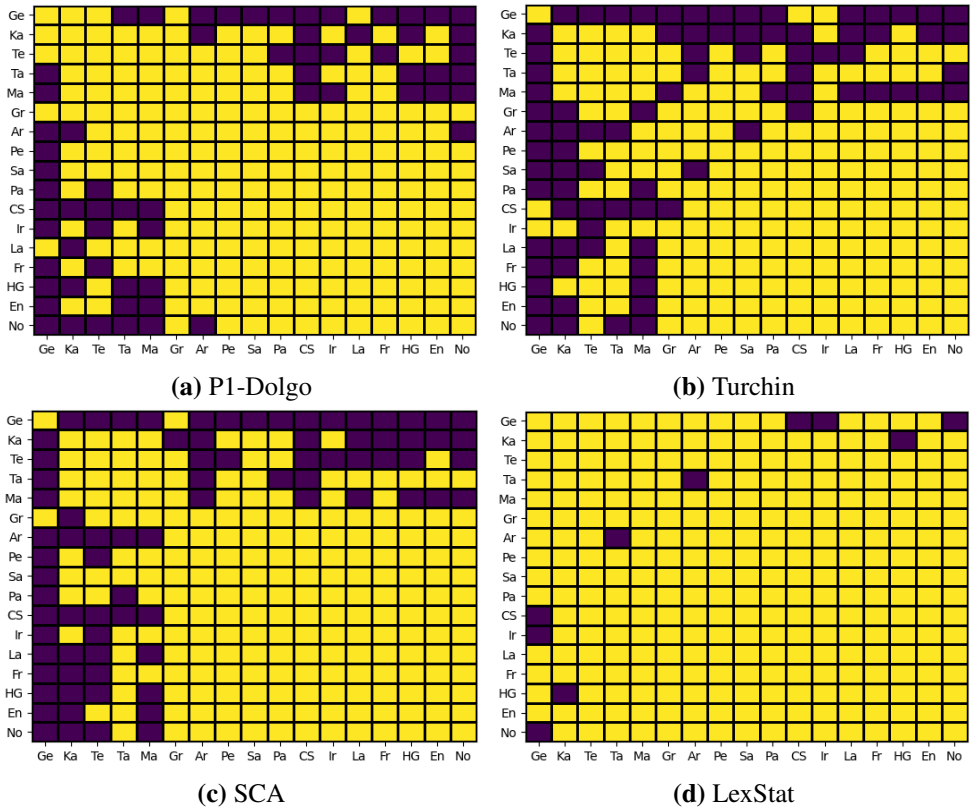
Table 6.6: Results of evaluation of macro families. Parentheses contain p-values.

6.6 Evaluation of Macro Families

We apply the tests on groupings of a few families from proposed macro families, namely Nostratic, Macro-Mayan, and Amerind. Under Nostratic, we test for groupings Dravidian-Indo-European (*Drav-IE*) and Dravidian-Indo-European-Kartvelian (*Drav-IE-Kart*) while we test Mayan-Mixe-Zoque (*May-MZ*) under Macro-Mayan and Mayan-Uto-Aztecan (*May-UAz*), Mayan-Mixe-Zoque-Uto-Aztecan (*May-MZ-UAz*) under Amerind. The results are tabulated in Table 6.6. While going by the p-values, the LRT test seems to support all of the mentioned families. However, the mean LRT statistic $\hat{\delta}$ is weak (negative or close to 0) for Drav-IE-Kart (Nostratic) and May-UAz, May-MZ-UAz (Amerind). In other words, by looking at Eq. (6.4), the alternate hypothesis H_a, i.e., having higher invariant sites is not preferred. Thus, it may be concluded that LRT is a highly sensitive test since the mere addition of a single language (Georgian) to a strongly supported group of 16 languages (Drav-IE) alters the outcome drastically. This is a desirable property since the presence of even a single anomaly, an unrelated language in this case, can be detected. Note that other combinations in Nostratic such as Drav-Kart or IE-Kart are much weaker and not well supported by the permutation test itself, which is elaborated as follows.

6.6.1 Analysis of Permutation tests on Nostratic

Bilateral significances on Nostratic grouping Drav-IE-Kart for various distance metrics are reported in Figure 6.4, where the pairwise relationships based on p-value (with threshold



(a) P1-Dolgo **(b)** Turchin

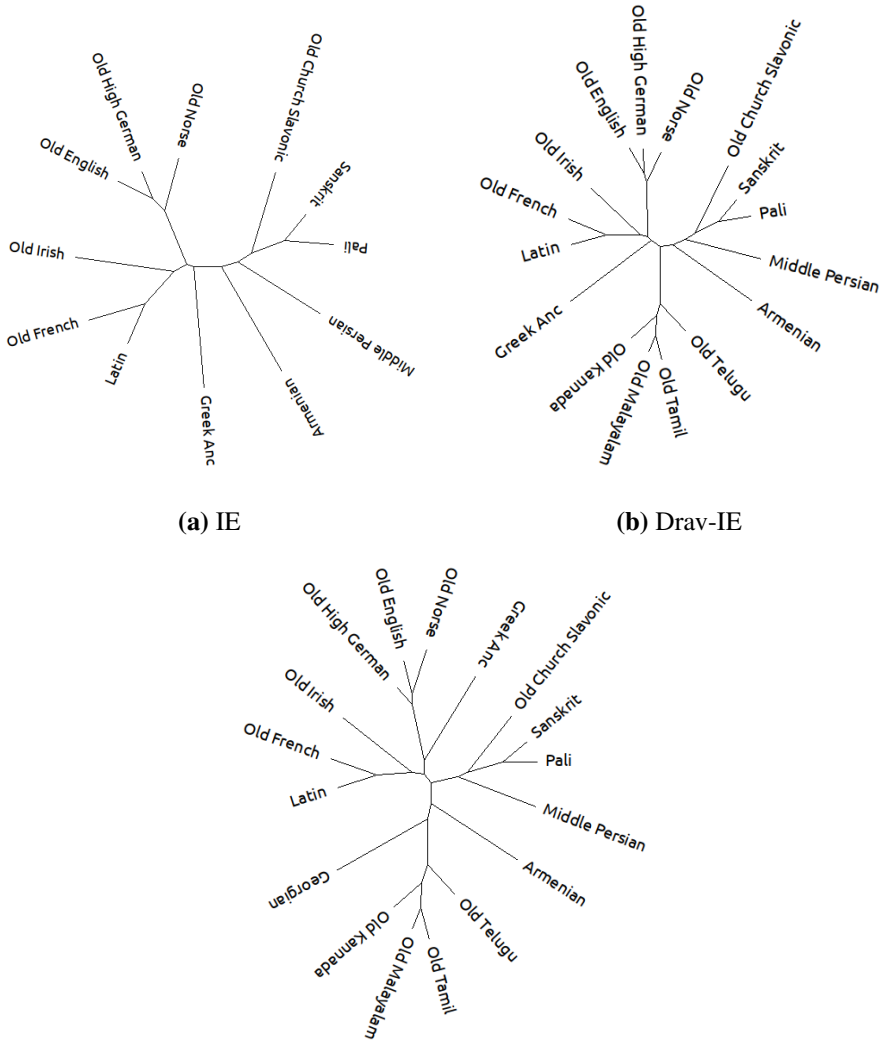
(c) SCA **(d)** LexStat

Figure 6.4: Bilateral (pairwise) significance among the languages of Nostratic grouping. The yellow shade implies that the relationship is statistically significant ($p < 0.05$), while the purple shade implies otherwise.

0.05) are color-coded. The computation follows the same steps as defined in §6.4.2 except that distances and similarities are calculated over pairs of languages instead of language clusters. This indeed forms the first iteration of a complete multilateral test.

The languages are abbreviated in Figure 6.4 as follows: Old Georgian (Ge), Old Kannada (Ka), Old Telugu (Te), Old Tamil (Ta), Old Malayalam (Ma), Ancient Greek (Gr), Old Armenian (Ar), Middle Persian (Pe), Sanskrit (Sa), Pali (Pa), Old Church Slavonic (CS), Old Irish (Ir), Latin (La), Old French (Fr), Old High German (HG), Old English (En) and Old Norse (No).

It is visible that for each metric, languages of the same family (IE and Drav) are almost always related pairwise. Secondly, many pairs from Drav-IE appear related. On the other hand, except for LexStat, Georgian shows to be related to at most two languages from the Drav-IE grouping. Yet, in the permutation tests for these metrics, except for Turchin (Table 6.6), Drav-IE-Kart appears significantly related with sometimes even good similarity scores (in the case of P1-Dolgo). All that can be concluded here is that, except for the



(a) IE **(b)** Drav-IE

(c) Drav-IE-Kart

Figure 6.5: Comparison of unrooted ML-trees on various groupings of Nostratic language families

LexStat metric, permutation tests are very sensitive to pairwise language comparisons and may not yield false positives. However, if Drav-IE-Kart is to be considered a valid grouping, these tests may be said to yield false negatives.

6.6.2 Analysis of ML-trees of Nostratic

Unrooted maximum likelihood trees (ML-trees) are drawn in Figure 6.5 on various sub-groupings of Nostratic using MEGA11 assuming the Poisson+I model. For the IE tree (Figure 6.5(a)), the sub-families, except for the position of Old Church Slavonic, are highly faithful reflecting the existing notions. For instance, the topology of the Germanic family,

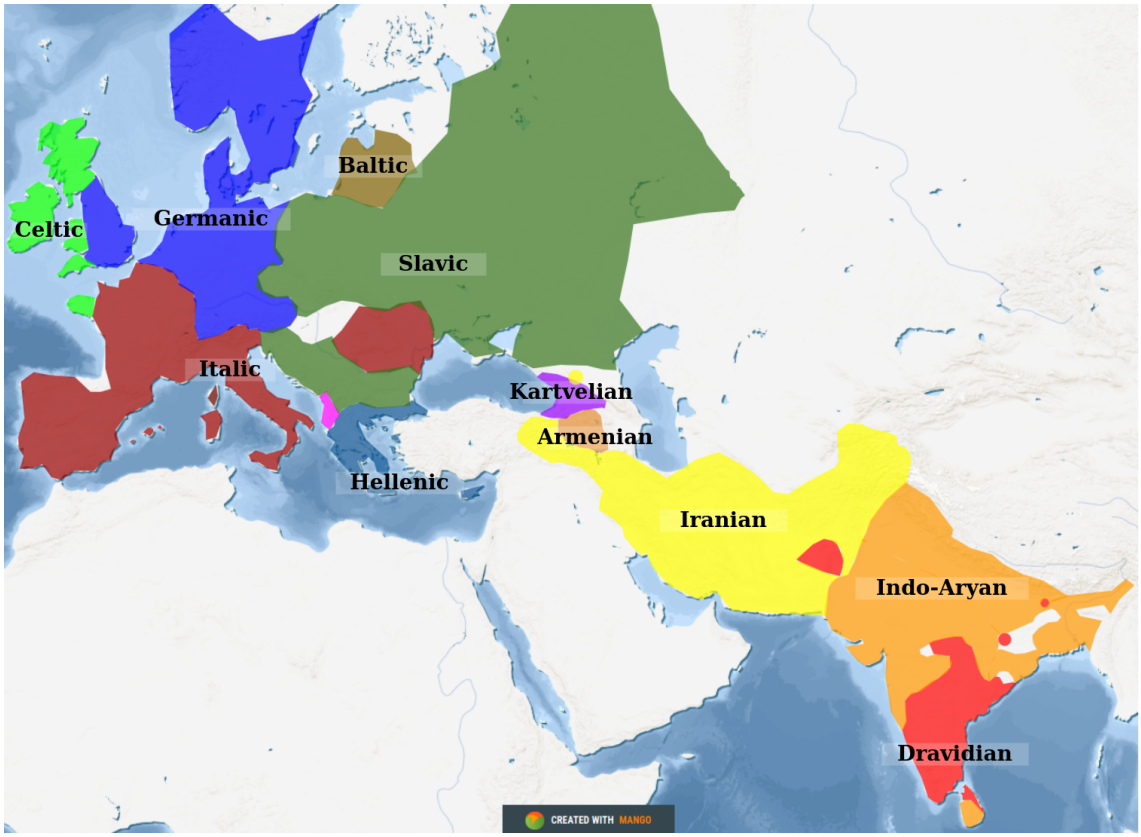


Figure 6.6: Geographical distribution of the language families/sub-families considered within the Nostratic grouping roughly around 10th century CE. (Created with *Mango*)

i.e., (Old Norse, (Old English, Old High German)) contains the valid West-Germanic branch (Old English, Old High German). Similarly, the Italo-Celtic group (Old Irish, (Latin, Old French)) is visible. Also, one can distinguish a clear boundary between Western and Eastern IE languages reflecting the geographical distribution. Only the position of Old Church Slavonic which appears intruded into the Indo-Iranian branch is problematic in this tree.

Furthermore, the addition of the Dravidian family in Drav-IE does not alter the IE topology (Figure 6.5(b)). It is intriguing to note the western inclination of Dravidian given its eastern geographical location in the present day. However, this is in line with the observation of Caldwell (1875), the founder of comparative Dravidian linguistics himself. This can be inferred by noticing few words of similar appearence across the two families: 'all' Ta, Ma, Ka, Te /elːa/ vs HG /al/, En /æɑɬ/, No /alːɪr/; 'many' Ta, Ma, Ka, Te /palV/ vs Gr /polýs/; 'one' Ta /onru/, Ma /onːə/ vs La /uːnum/, Ir /oin/, HG /aɪn/; 'short' Ta /kuru/, Ma /kuri/ vs OH /skurt/, Fr /kort/; 'woman' Ta, Ma, Ka /peɳ/ vs Ir /ben/, HG /kwɛn/; 'root'

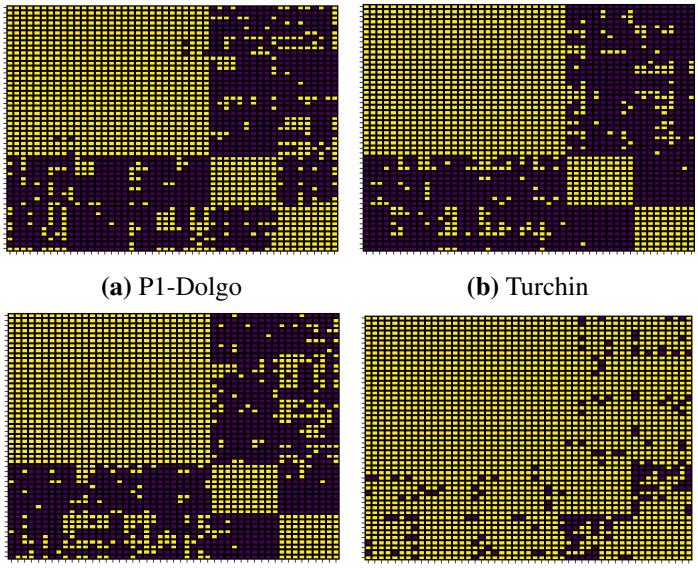
Ta, Ma, Te /ʋeːɾ(u)/ vs HG /wurts/, En /wyrt/; ‘blood’ Ta, Ma /kuɾuṯ̪i/ vs Ir /kruː/, CS /krʊvɪ/; ‘horn’ Ta /koːʈu/ Ka, Te /kombu/ vs La /kornu/, Gr /kéras/, and several others. Most of these similarities were already pointed out by Caldwell (1875), who also identifies them as similarities among Dravidian and ‘Western’ Indo-European languages. While, there do exist potential cognates with the neighboring languages as found in ‘black’ Ta, Ma, Ka /karu/ vs Sa /kr̥ʂɳa/ or in ‘fruit’ Ta, Ma /paʟam/ vs Sa /p^{h}alam/, those with the far away languages are clearly the ones that are binding the two families. This fact is also reflected in the position of the Dravidian family in the tree topology. This would be an interesting scenario to research upon, which has unfortunately not received due attention so far in the community of historical linguists.

Futhermore, the addition of Georgian invalidates the West-Germanic branch as well as pushes Old Greek problematically into the Western group (Figure 6.5(c)). However, much of the topology is undisturbed and one can also notice how the languages/families that are located south of the Caucasus namely, Armenian, Georgian, and Dravidian are grouped. For the geographical distributions of these languages, see Figure 6.6. Overall, it may be concluded that the addition of unrelated or weakly related languages can alter the actual topology.

Similar analyses in case of Macro-Mayan and Amerind families are provided in Fig. 6.7 & 6.8, where one can observe similar perturbations in topology (see Fig. 6.8) of one family (Mayan) in presence of others (Mixe-Zoque and Uto-Aztecan).

6.7 Summary

In this chapter, we have presented a likelihood ratio test based on the proportions of invariant sites to determine the genetic relatedness of a group of languages. Our proposed test does not yield false positives, which is in contrast with previous permutation-based tests that proved to be good only for pairwise language comparisons and not for validating a language group. By applying this test, we have found strong supporting evidence for macro-families such as Dravidian-Indo-European, Macro-Mayan (for Mayan-Mixe-Zoque, and weak evidence for Nostratic (Dravidian-Indo-European-Kartvelian) and Amerind (for



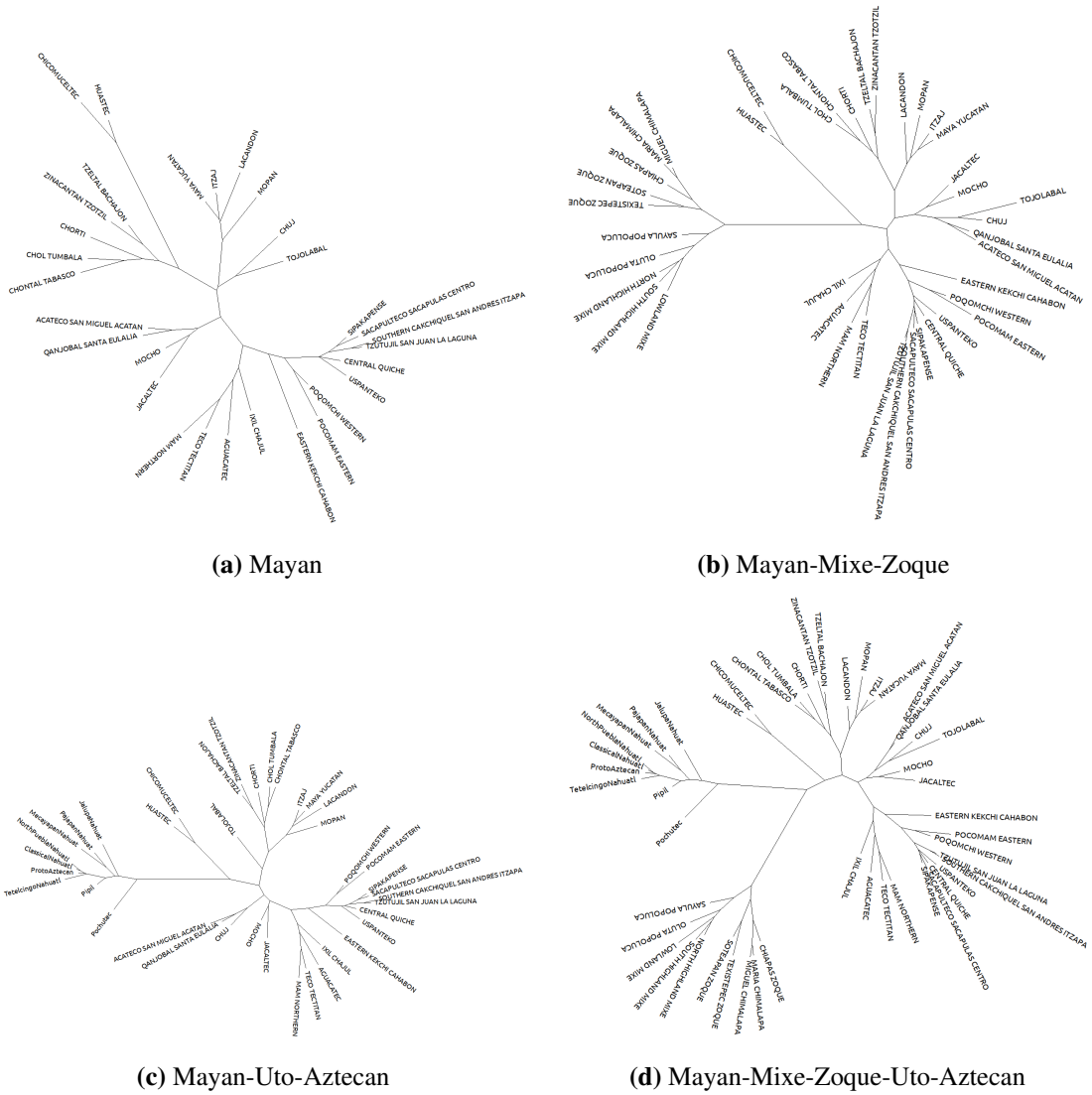
(a) P1-Dolgo **(b)** Turchin

(c) SCA **(d)** LexStat

Figure 6.7: Bilateral (pairwise) significance among the languages of Macro-Mayan/Amerind grouping. The yellow shade implies that the relationship is statistically significant ($p < 0.05$), while the purple shade implies otherwise. While moving across the diagonal, the first cluster of significantly related languages is that of Mayan, the second is that of Mixe-Zoque and the thrid, Uto-Aztecan

Mayan-Uto-Aztecan). Through secondary analyses, we have also shown that probabilistic-based methods are superior to distance-based ones based on tree construction and the correlation of topologies with geography. In this work we did not touch upon semantic shifts, i.e., words changing meaning over time; for example, the word *quick* initially meant 'lively'. While considering semantic shifts may provide room for data manipulation favoring any particular hypothesis, few semantic slots such as 'bark'-'skin' are often found to have common words. In such cases, the slots may be merged into one as suggested by Kessler (2001). Further, incorporating grammatical features may improve the constructed trees. Nevertheless, these are not included in significant testing since morphological reconstruction comes later in the comparative method while significant tests focus around the first step i.e., they question if gathered cognates are chance occurences.

In summary, before constructing phylogenies of a group of languages, the relatedness of the group should be established through a significance test such as the one we have presented. Otherwise, the phylogenic grouping would not only be questionable but may also alter the topology of a related sub-group.



(a) Mayan **(b)** Mayan-Mixe-Zoque

(c) Mayan-Uto-Aztecan **(d)** Mayan-Mixe-Zoque-Uto-Aztecan

Figure 6.8: Comparison of unrooted ML-trees on various groupings of Macro-Mayan/Amerind language families

Limitations

The values of P^0_{inv} and P^a_{inv} (§6.3.5) are roughly decided based on the estimated ones from two examples, namely, Afrasian-Lolo-Burmese as a negative example and Indo-European as a positive example. The question of what should be the most appropriate values that should make the test optimal is not addressed here. Ideally, to address this question, more data is needed with several positive and negative examples to search for optimal values of these parameters. Also, the exact values may require calibration according to the phylogenetic software used since there could be significant differences in the implementations. Secondly, while analyzing Nostratic languages, Uralic, an important

language family, has not been included due to the selection criteria (§6.4.1) that the languages should have been attested before 10th century CE. To include Uralic, the (Nostratic) languages that are attested around the same period as the earliest attested ones from Uralic (roughly 1300 CE onwards) should be considered to make ‘fair’ comparisons.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The thesis has started by mentioning three central problems of computational historical linguistics namely proto-language reconstruction, cognate detection, and significance testing of the genetic relatedness, along with the existing methods and their limitations. Certain aims have been then set in Chapter 3 which are listed as follows.

- To develop computational tools for the aforementioned three problems that can overcome the limitations of existing methods.
- To maintain an integration of biological insights throughout following the tradition of computational historical linguistics.
- To evaluate the performances of the proposed methods against those of existing methods especially at low resource settings.
- To maintain multilingualism, i.e., to ensure language diversity through out the experiments.
- To bear broader implications, especially by applying the significance tests on proposed macro-families.

The contributions of this thesis fulfilling the above aims are summarized as follows.

7.1.1 Development of Computational Tools

To overcome the limitations of the existing methods, the following methods have been developed for the problems addressed in this thesis:

1. We have proposed Cognate Transformer (CogTran) for the problems of proto-language reconstruction and cognate reflex prediction which could outperform the existing methods, especially when pre-trained, demonstrating the efficacy of transfer learning.

2. We have proposed CogTran2, i.e., CogTran with link prediction head for the problem of automatic cognate detection which could outperform the existing methods in the presence of little additional supervision. Thus, the model could utilize well the labeled information, unlike the previous methods.

3. We proposed a likelihood ratio test of genetic relationship based on the proportions of the phonetically conserved sites which, unlike the existing methods, does not exhibit the problems of false positives nor does it require reconstructed proto-forms.

7.1.2 Integration of Biological Insights

Following the tradition of computational historical linguistics, the methods proposed in this thesis maintain a constant integration of insights from evolutionary biology in the following ways:

1. CogTran is based on a protein language model, namely MSA transformer (Rao et al., 2021) that acts on a multiple sequence alignment (MSA) input.

2. CogTran2 incorporates outer-product mean and pairwise module from the protein structure predictor AlphaFold2 (Jumper et al., 2021), among which pairwise module was supposed to capture transitivity of cognacy. Further, it is an end-to-end architecture that directly takes an MSA as input and gives cognate cluster linkages as output. This workflow contributes to a significant speed-up since it circumvents pairwise computations usually found in the previous methods.

3. LRT is inspired by the phylogenetic hypothesis testing approach found in molecular phylogenetics.

7.1.3 Performance Evaluation

The performances of all the methods proposed in this thesis have been evaluated against those of previous methods and are found to be better than them. In particular, CogTran proves to be efficient at higher test proportions i.e., at low resource settings. Similarly, CogTran2 could be adapted to new language families by only learning from a few concepts. Thus, these models are useful in low-resource settings which reflect the linguistic reality.

7.1.4 Multilinguality

All the datasets in this thesis come from diverse language families, thus fulfilling the criteria of multilingualism in our experiments. The results give confidence that the models can be easily applied to any language family.

7.1.5 Broader Implications

Finally, this thesis does not just provide better assistance to the historical linguists but also bears some important results. Application of significance tests on macro-families (§6.6) namely Macro-Mayan and Nostratic respectively suggest genetic relatedness of Mayan-Mixe-Zoquen and Indo-European-Dravidian. The potential relatedness of Indo-European and Dravidian has important consequences of rethinking the past of Indian subcontinent where these languages are native to about 98%[1] of the population. Further, it can also potentially impact the current understanding of the homelands of these language families.

7.2 Future Work

Several problems can be looked upon in the future related to either those emerging out of the limitations of the proposed methods or those unaddressed in this thesis. Some of these

[1]Indo-Aryan consists of about 78%, while Dravidian 20% (https://www.britannica.com/topic/Indo-Aryan-languages)

are described as follows.

7.2.1 Automated Sound Correspondence Inference

We briefly mentioned some aspects of sound correspondences in §4.5.5 and §5.6.3, especially in the latter case where CogTran2 appeared to have learned some sound correspondences such as Ancient Greek $k \sim$ German h. However, the automatic extraction of such sound correspondences from the wordlists of related languages is not mentioned. A recent approach to this problem by List (2019a) solves it by framing the task as a clique-cover problem. However, conditional sound changes, such as palatalization, are not addressed therein which otherwise seem to be captured by CogTran (§4.5.5). Overall, it would be interesting for future work to extract sound correspondences from CogTran or CogTran2 and compare them with that of the clique-cover-based method.

7.2.2 Multiple Sequence Alignment as part of the Neural Pipeline

Algorithms employed to build MSAs in this thesis are either from the LingPy library (List and Forkel, 2021) or Clustal W (Larkin et al., 2007). LingPy internally uses DIALIGN (Morgenstern et al., 1998), a progressive alignment tool which considers both global and local alignments. Another tool used in building MSA from phonetic sequences is T-coffee (Notredame et al., 2000) used in Jäger (2022). All three tools namely, T-coffee, DIALIGN, and Clustal W are based on progressive alignment. Other popular MSA tools in bioinformatics include progressive alignment-based MUSCLE (Edgar, 2004), fast-fourier transform-based MAFFT (Katoh et al., 2002), and profile Hidden Markov Model-based Clustal Omega (Sievers and Higgins, 2014).

Since all the methods in this thesis start from MSA, the MSA quality becomes a bottleneck in their performances. To bring further improvements, it is sensible to include MSA in the neural pipelines at least allowing for minor differentiable changes on top of the MSAs produced by the tools. Recent neural approaches to MSA in bioinformatics involve reinforcement learning (Mircea et al., 2018; Ramakrishnan et al., 2018; Jafari et al., 2019), transformer (Dotan et al., 2023), or both (Liu et al., 2023).

7.2.3 Phoneme Substitution Models in Phylogenetic Inference

The substitution model employed to compute likelihoods in Chapter 6 (§6.3.2) assumes equal rates of character substitution. However, in reality, sound change rates are not identical among various sounds. For instance, it is more likely for an alveolar stop *T* to alternate with a sibilant *S* or a palatal stop *C* than with a velar stop *K*. In molecular biology, such a situation is handled by models with unequal mutation rates such as that of Kimura (1980), which assumes unequal mutation rates between a purine (A or G) and a pyrimidine (C or T). In the case of amino acids, the rates are often determined empirically such as in the case of BLOSUM62 matrix (Henikoff and Henikoff, 1992). Similar transition matrices are found in the computation of sound class-based alignment of List (2010), which can be incorporated into the likelihood computation.

Further, the model can be enriched by imposing regularity of sound changes and non-uniform rates across different prosodic positions such as word-initial versus medial. Further, LRT assumes site independence (§6.3.3) in the likelihood computation. More complex models may be proposed lifting this assumption. However, the number of parameters should be restricted to small numbers considering the amount of data available.

Other potential works can focus on addressing the limitations mentioned in previous chapters such as the problem of metathesis (§4.5.3), partial cognacy (§5.6.3), and further inclusion of languages and families in the significance testing mentioned in Chapter 6.

7.3 Concluding Remarks

Thus, this thesis has presented a cognate transformer and a likelihood ratio test, both inspired by biological sequence evolution, for addressing three central problems in the study of language change namely gathering cognates, reconstructing ancestral words, and ascertaining genetic relationships, bearing substantial results. One salient outcome is major statistical support for the Indo-European affiliation of the Dravidian languages.

Through this thesis, I learned how elegant solutions can be provided by observing cross-disciplinarily for structurally similar problems. I hope this thesis will generate a common

interest in the readers toward historical linguistics from a computational perspective as well as a general sense of appreciation for the languages spoken by one's ancestors.

References

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2023a. Cognate transformer for automated phonological reconstruction and cognate reflex prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6852–6862, Singapore. Association for Computational Linguistics.

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2023b. Creation of a digital Rig Vedic index (Anukramani) for computational linguistic tasks. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 89–96, Canberra, Australia (Online mode). Association for Computational Linguistics.

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2024a. Automated Cognate Detection as a Supervised Link Prediction Task with Cognate Transformer. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 965–975, St. Julian's, Malta. Association for Computational Linguistics.

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2024b. A likelihood ratio test of genetic relationship among languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2559–2570, Mexico City, Mexico. Association for Computational Linguistics.

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

Maria Anisimova and Olivier Gascuel. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, 55(4):539–552.

Quentin D Atkinson and Russell D Gray. 2005. Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic biology*, 54(4):513–526.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Aditya Bhargava and Grzegorz Kondrak. 2009. Multiple word alignment with Profile Hidden Markov Models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, pages 43–48, Boulder, Colorado. Association for Computational Linguistics.

M J Bishop and A E Friday. 1987. Tetrapod relationships: The molecular evidence. *Molecules and morphology in evolution: Conflict or compromise*, pages 123–139.

Frederic Blum and Johann-Mattis List. 2023. Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 52–64, Dubrovnik, Croatia. Association for Computational Linguistics.

Allan R Bomhard and John C Kerns. 1994. *The Nostratic macrofamily: A study in distant linguistic relationship*. De Gruyter Mouton.

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.

Cecil H Brown, Eric W Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals*, 61(4):285–308.

Robert Caldwell. 1875. *A comparative grammar of the Dravidian or South-Indian family of languages*. Trübner.

Lyle Campbell. 1997. *American Indian languages: The historical linguistics of Native America*, volume 4. Oxford University Press, USA.

Lyle Campbell. 2013. *Historical linguistics*. Edinburgh University Press.

Giuseppe G. A. Celano. 2022. A transformer architecture for the prediction of cognate reflexes. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 80–85, Seattle, Washington. Association for Computational Linguistics.

Alina Maria Ciobanu and Liviu P. Dinu. 2018. Ab initio: Automatic Latin proto-word reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1604–1614, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Michael A. Covington. 1996. An Algorithm to Align Words for Historical Comparison. *Computational Linguistics*, 22(4):481–496.

Johannes Dellert. 2018. Combining information-weighted sequence alignment and sound correspondence models for improved cognate detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3123–3133, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rebecca W Doerge and GA1206957 Churchill. 1996. Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1):285–294.

Edo Dotan, Yonatan Belinkov, Oren Avram, Elya Wygoda, Noa Ecker, Michael Alburquerque, Omri Keren, Gil Loewenthal, and Tal Pupko. 2023. Multiple sequence alignment as a sequence-to-sequence learning problem. In *International Conference on Learning Representations*.

Robert C. Edgar. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Daniel P Faith. 1991. Cladistic permutation tests for monophyly and nonmonophyly. *Systematic Zoology*, 40(3):366–375.

Joseph Felsenstein. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249.

Joseph Felsenstein. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376.

Nick Goldman, Jon P Anderson, and Allen G Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*, 49(4):652–670.

Joseph H Greenberg. 1963. The languages of Africa. *International Journal of American Linguistics*.

Joseph H Greenberg. 1971. The Indo-Pacific hypothesis. *Current Trends in Linguistics*, 8:807–871.

Joseph H Greenberg. 1987. *Language in the Americas*. Stanford University Press.

Joseph H Greenberg. 2000. *Indo-European and its closest relatives: The Eurasiatic language family, volume 1, grammar*, volume 1. Stanford University Press.

Joseph H Greenberg. 2005. *Genetic linguistics: Essays on theory and method*. OUP Oxford.

Steven Henikoff and Jorja G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89 22:10915–9.

John P Huelsenbeck and JJ Bull. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. *Systematic Biology*, 45(1):92–98.

John P Huelsenbeck, David M Hillis, and Rasmus Nielsen. 1996. A likelihood-ratio test of monophyly. *Systematic Biology*, 45(4):546–558.

John P Huelsenbeck and Fredrik Ronquist. 2001. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Reza Jafari, Mohammad Masoud Javidi, and Marjan Kuchaki Rafsanjani. 2019. Using deep reinforcement learning approach for solving the multiple sequence alignment problem. *SN Applied Sciences*, 1.

Gerhard Jäger. 2015. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, 112(41):12752–12757.

Gerhard Jäger. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5(1).

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

Gerhard Jäger. 2022. Bayesian phylogenetic cognate prediction. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 63–69, Seattle, Washington. Association for Computational Linguistics.

Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216, Valencia, Spain. Association for Computational Linguistics.

Gerhard Jäger and Pavel Sofroniev. 2016. Automatic cognate classification with a support vector machine. In *Proceedings of the 13th conference on Natural Language Processing*, volume 16, pages 128–134. RUB Bochum.

William Jones. 1824. *Discourses delivered before the Asiatic Society: and miscellaneous papers, on the religion, poetry, literature, etc., of the nations of India*. CS Arnold.

Thomas H Jukes, Charles R Cantor, et al. 1969. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Shaun M Kandathil, Joe G Greener, Andy M Lau, and David T Jones. 2022. Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins. *Proceedings of the National Academy of Sciences*, 119(4):e2113348119.

Diptesh Kanojia, Raj Dabre, Shubham Dewangan, Pushpak Bhattacharyya, Gholamreza Haffari, and Malhar Kulkarni. 2020. Harnessing cross-lingual features to improve cognate detection for low-resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1384–1395, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Diptesh Kanojia, Prashant Sharma, Sayali Ghodekar, Pushpak Bhattacharyya, Gholamreza Haffari, and Malhar Kulkarni. 2021. Cognition-aware cognate detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3281–3292, Online. Association for Computational Linguistics.

Alexei Kassian, Mikhail Zhivlov, and George Starostin. 2015. Proto-Indo-European-Uralic comparison from the probabilistic point of view. *Journal of Indo-European Studies*, 43(3-4):301–347.

Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066.

Patrick J Keeling and Jeffrey D Palmer. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618.

Brett Kessler. 2001. The significance of word lists. *Stanford*.

Brett Kessler. 2007. Word Similarity Metrics and Multilateral Comparison. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 6–14, Prague, Czech Republic. Association for Computational Linguistics.

Brett Kessler. 2008. The Mathematical Assessment of Long-Range Linguistic Relationships. *Language and Linguistics Compass*, 2(5):821–839.

Brett Kessler. 2015. Response to Kassian et al., Proto-Indo-European-Uralic comparison from the probabilistic point of view. *Journal of Indo-European Studies*, 43(3-4):357–367.

Brett Kessler and Annukka Lehtonen. 2006. Multilateral comparison and significance testing of the Indo-Uralic question. *Phylogenetic methods and the prehistory of languages*, pages 33–42.

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.

Motoo Kimura. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120.

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

Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 70–79, Seattle, Washington. Association for Computational Linguistics.

Grzegorz Kondrak. 1999. Alignment of phonetic sequences. *Department of Computer Science, University of Toronto, Tech. Rep. CSRG-402*.

Grzegorz Kondrak. 2000. A New Algorithm for the Alignment of Phonetic Sequences. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Mark A Larkin, Gordon Blackshields, Nigel P Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948.

Vladimir I Levenshtein. 1965. Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov. *Doklady Akademij Nauk SSSR*, 163(4):845–848.

Stephan Hillyer Levitt. 1998. Is there a genetic relationship between indo-european and dravidian? *Journal of Indo-European studies*, 26(1/2):131.

Dylan Lewis, Winston Wu, Arya D. McCarthy, and David Yarowsky. 2020. Neural transduction for multilingual lexical translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4373–4384, Barcelona, Spain (Online). International Committee on Computational Linguistics.

JM List and R Forkel. 2022. Lingrex. *Linguistic reconstruction with LingPy*.

Johann-Mattis List. 2010. Sca: Phonetic alignment based on sound classes. In *European Summer School in Logic, Language and Information*, pages 32–51. Springer.

Johann-Mattis List. 2012a. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France. Association for Computational Linguistics.

Johann-Mattis List. 2012b. Sca: Phonetic alignment based on sound classes. In *New Directions in Logic, Language and Computation: ESSLLI 2010 and ESSLLI 2011 Student Sessions. Selected Papers*, pages 32–51. Springer.

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):247–258.

Johann-Mattis List and Robert Forkel. 2021. Lingpy. a python library for historical linguistics. version 2.6.9.

Johann-Mattis List, Robert Forkel, and Nathan Hill. 2022a. 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, Simon J Greenhill, and Russell D Gray. 2017. The potential of automatic word comparison for historical linguistics. *PloS one*, 12(1):e0170046.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Berlin, Germany. Association for Computational Linguistics.

Johann-Mattis List, Ekaterina Vylomova, Robert Forkel, Nathan Hill, and Ryan Cotterell. 2022b. 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.

Yuhang Liu, Hao Yuan, Qiang Zhang, Zixuan Wang, Shuwen Xiong, Naifeng Wen, and Yongqing Zhang. 2023. Multiple sequence alignment based on deep reinforcement learning with self-attention and positional encoding. *Bioinformatics*, 39(11):btad636.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Nhan Ly-Trong, Suha Naser-Khdour, Robert Lanfear, and Bui Quang Minh. 2022. AliSim: a fast and versatile phylogenetic sequence simulator for the genomic era. *Molecular Biology and Evolution*, 39(5):msac092.

Wesley Mackay and Grzegorz Kondrak. 2005. Computing word similarity and identifying cognates with pair hidden Markov models. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 40–47, Ann Arbor, Michigan. Association for Computational Linguistics.

Roddy MacSween and Andrew Caines. 2020. An expectation maximisation algorithm for automated cognate detection. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 476–485, Online. Association for Computational Linguistics.

Thomas Mailund and Christian NS Pedersen. 2004. QDist—Quartet distance between evolutionary trees. *Bioinformatics*, 20(10):1636–1637.

James P Mallory and Douglas Q Adams. 2006. *The Oxford introduction to proto-Indo-European and the proto-Indo-European world*. Oxford University Press, USA.

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.

Claudio Mirabello and Björn Wallner. 2019. rawmsa: End-to-end deep learning using raw multiple sequence alignments. *PloS one*, 14(8):e0220182.

Ioan-Gabriel Mircea, Iuliana Bocicor, and Gabriela Czibula. 2018. A reinforcement learning based approach to multiple sequence alignment. In *Soft Computing Applications*, pages 54–70, Cham. Springer International Publishing.

B Morgenstern, K Frech, A Dress, and T Werner. 1998. DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, 14(3):290–294.

Abhijnan Nath, Rahul Ghosh, and Nikhil Krishnaswamy. 2022. Phonetic, semantic, and articulatory features in Assamese-Bengali cognate detection. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 41–53, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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.

Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2020. Multi-source neural machine translation with missing data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:569–580.

Cédric Notredame, Desmond G Higgins, and Jaap Heringa. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment11edited by j. thornton. *Journal of Molecular Biology*, 302(1):205–217.

Robert L Oswalt. 1970. The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior*, 3(3):117–129.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS One*, 6(6):e20109.

William Poser and Lyle Campbell. 2008. Language Classification: History and Methods.

Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027, Osaka, Japan. The COLING 2016 Organizing Committee.

Taraka Rama. 2018. Similarity dependent Chinese restaurant process for cognate identification in multilingual wordlists. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 271–281, Brussels, Belgium. Association for Computational Linguistics.

Taraka Rama and Johann-Mattis List. 2019. An automated framework for fast cognate detection and Bayesian phylogenetic inference in computational historical linguistics. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6225–6235, Florence, Italy. Association for Computational Linguistics.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400, New Orleans, Louisiana. Association for Computational Linguistics.

Ramchalam Kinattinkara Ramakrishnan, Jaspal Singh, and Mathieu Blanchette. 2018. Rlalign: A reinforcement learning approach for multiple sequence alignment. In *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 61–66.

Priya Rani, Koustava Goswami, Adrian Doyle, Theodorus Fransen, Bernardo Stearns, and John P. McCrae. 2023. Findings of the SIGTYP 2023 shared task on cognate and derivative detection for low-resourced languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 126–131, Dubrovnik, Croatia. Association for Computational Linguistics.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. 2021. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR.

Donald A Ringe. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society*, 82(1):1–110.

Donald A Ringe. 1996. The mathematics of 'Amerind'. *Diachronica*, 13(1):135–154.

Donald A Ringe and Joseph F Eska. 2013. *Historical linguistics: Toward a twenty-first century reintegration*. Cambridge University Press.

Fredrik Ronquist, Maxim Teslenko, Paul Van Der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. 2012. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542.

Fabian Sievers and Desmond G. Higgins. 2014. *Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences*, pages 105–116. Humana Press, Totowa, NJ.

Temple F Smith and Michael S Waterman. 1981. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.

Robert R Sokal and Charles D Michener. 1975. A statistical method for evaluating systematic relationships. *Multivariate statistical methods, among-groups covariation*, page 269.

Robert R. Sokal and Charles Duncan Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas science bulletin*, 38:1409–1438.

PS Subrahmanyam. 2011. The prakrit grammatians: Historical linguists of ancient india. *Indian Linguistics*, 72(1-4):230.

Koichiro Tamura, Glen Stecher, and Sudhir Kumar. 2021. MEGA11: molecular evolutionary genetics analysis version 11. *Molecular Biology and Evolution*, 38(7):3022–3027.

Julie D Thompson, Toby J Gibson, and Des G Higgins. 2003. Multiple sequence alignment using clustalw and clustalx. *Current protocols in bioinformatics*, (1):2–3.

Peter Turchin, Ilia Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship*, (5 (48)):117–126.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Edward Orlando Wiley and Bruce S Lieberman. 2011. *Phylogenetics: Theory and practice of phylogenetic systematics*. John Wiley & Sons.

S. S. Wilks. 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Appendix A

Row and Column Attentions

The details of row and column attentions of Cognate Transformer are provided here, which are based on MSA Transformer (Rao et al., 2021).

Given the MSA embedding $\boldsymbol{m} \in \mathbb{R}^{r \times c \times d}$ with r rows (number of sequences), c columns (sequence length), and embedding dimension d, the row attention output for a single head is computed as follows,

$$\boldsymbol{o}_{ij} \leftarrow \sum_{k} \text{softmax}_k \left(\frac{1}{\sqrt{c}} \boldsymbol{q}_{ij}^{\top} \boldsymbol{k}_{ik} \right) \boldsymbol{v}_{ik}$$

Queries $\boldsymbol{q}_{ij}$, keys $\boldsymbol{k}_{ik}$ and values $\boldsymbol{v}_{ik}$ are linear projections of $\boldsymbol{m}$ and have the usual connotations. In other words,

$$\boldsymbol{q}_{ij}, \boldsymbol{k}_{ij}, \boldsymbol{v}_{ij} \leftarrow \text{LinearNoBias}(\boldsymbol{m}_{ij})$$

Similarly, column attention computation for a single head is as follows,

$$\boldsymbol{o}_{ij} \leftarrow \sum_{k} \text{softmax}_k \left(\frac{1}{\sqrt{c}} \boldsymbol{q}_{ij}^{\top} \boldsymbol{k}_{kj} \right) \boldsymbol{v}_{kj}$$

This is followed by the usual concatenation and projection of the representations from different heads. In other words, let $\boldsymbol{o}_{ij}^{h}$ be attention output from head h,

$$\boldsymbol{m}'_{ij} \leftarrow \text{Linear}(\text{concat}_h \boldsymbol{o}_{ij}^{h})$$

All initial linear layers project to intermediate size, while projection at the end brings it back to the hidden size d (§5.4). Implementation of CogTran uses equal hidden and intermediate sizes (§5.5.2). The row attention is tied, i.e., uses the same attention map across rows (Rao et al., 2021). On the other hand, column attention is untied, i.e., attention maps are different for each column. This would be useful since sound changes vary across positions in a word.

Appendix B

Miscellaneous Details of CogTran2

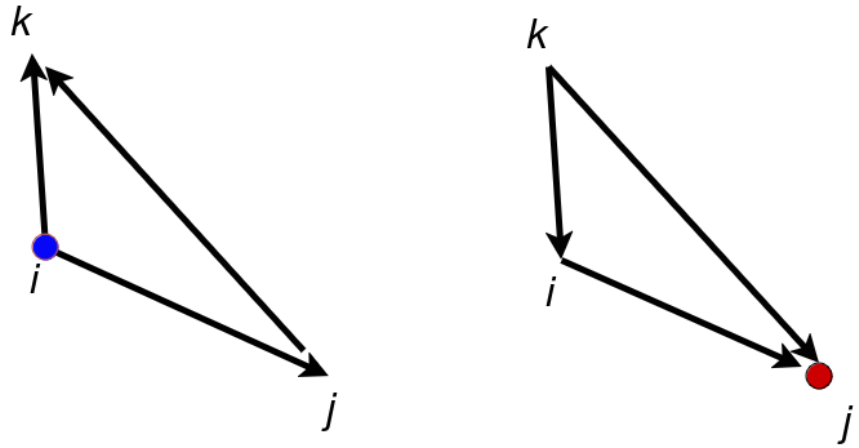


Figure B.1: Outgoing edges / around starting node i (left) and incoming edges / around ending node j (right)

Additional details of the modules used in CogTran2 (Chapter 5) are provided here, which are based on AlphaFold2 (Jumper et al., 2021).

B.1 Outer Product Mean

If $\boldsymbol{m} \in \mathbb{R}^{r \times c \times d}$ (§5.4.2) is the output of CogTran, outer product mean is performed as follows,

$$\boldsymbol{z}_{ij} \leftarrow \text{Linear}\left(\text{mean}_k \left\{\text{Linear}(\boldsymbol{m}_{ik}) \otimes \text{Linear}(\boldsymbol{m}_{jk})\right\}\right)$$

This operation is preceded by a layer norm. Note that this is slightly different than in

the case of AlphaFold2, since in our case, the outer product acts over rows while the mean over columns. The roles of rows and columns get interchanged in the case of AlphaFold2.

B.2 Triangle Multiplication Updates

The triangle multiplication update at of an edge representation $\boldsymbol{z}_{ij}$ of dimension d (§5.4.4), using the representations of outgoing edges is given by,

$$\boldsymbol{z}'_{ij} \leftarrow \sigma(\text{Linear}(\boldsymbol{z}_{ij})) \cdot \text{Linear}\left(\text{LayerNorm}\left(\textstyle\sum_k \sigma(\text{Linear}(\boldsymbol{z}_{ik})) \cdot \sigma(\text{Linear}(\boldsymbol{z}_{jk}))\right)\right)$$

In the case of using the incoming edges, it is performed similarly as follows,

$$\boldsymbol{z}'_{ij} \leftarrow \sigma(\text{Linear}(\boldsymbol{z}_{ij})) \cdot \text{Linear}\left(\text{LayerNorm}\left(\textstyle\sum_k \sigma(\text{Linear}(\boldsymbol{z}_{ki})) \cdot \sigma(\text{Linear}(\boldsymbol{z}_{kj}))\right)\right)$$

B.3 Triangle Attention

Attention computation for a single head around the starting node i of an edge ij is given as follows,

$$\boldsymbol{o}_{ij} \leftarrow \sum_k \text{softmax}_k \left(\frac{1}{\sqrt{c}} \boldsymbol{q}_{ij}^\top \boldsymbol{k}_{ik} + \boldsymbol{b}_{jk} \right) \boldsymbol{v}_{ik}$$

Queries $\boldsymbol{q}_{ij}$, keys $\boldsymbol{k}_{ik}$ and values $\boldsymbol{v}_{ik}$ are linear projections of the edge representations $\boldsymbol{z} \in \mathbb{R}^{r \times r \times d}$ (§5.4.4) and have the usual connotations. In other words,

$$\boldsymbol{q}_{ij}, \boldsymbol{k}_{ij}, \boldsymbol{v}_{ij} \leftarrow \text{LinearNoBias}(\boldsymbol{z}_{ij})$$

The bias $\boldsymbol{b}_{jk}$, another linear projection of the edge representation $\boldsymbol{z}_{jk}$, with projection weights, comes from the edge jk viz., opposite of node i. Similarly, attention computation for a single head around the ending node j is given as follows,

$$\boldsymbol{o}_{ij} \leftarrow \sum_k \text{softmax}_k \left(\frac{1}{\sqrt{c}} \boldsymbol{q}_{ij}^\top \boldsymbol{k}_{kj} + \boldsymbol{b}_{ki} \right) \boldsymbol{v}_{kj}$$

The pairwise modules use gated attention. In other words, given $\boldsymbol{o}_{ij}^h$ for attention head h, the output for this head would be,

$$\boldsymbol{o}_{ij}'^h \leftarrow \sigma(\text{Linear}(\boldsymbol{z}_{ij})) \cdot \boldsymbol{o}_{ij}^h$$

This is followed by the usual concatenation and projection of the representations from different heads:

$$\boldsymbol{z}'_{ij} \leftarrow \text{Linear}(\text{concat}_h \boldsymbol{o}_{ij}'^h)$$

All initial linear layers project to intermediate size, while projection at the end brings it back to the hidden size d (§5.4). Implementation of CogTran2 uses equal hidden and intermediate sizes (§5.5.2).

Appendix C

BCubed Cluster Evaluation Metrics

BCubed metrics for evaluating the quality of the clustering algorithms are outlined here as introduced by Amigó et al. (2009) based on Bagga and Baldwin (1998).

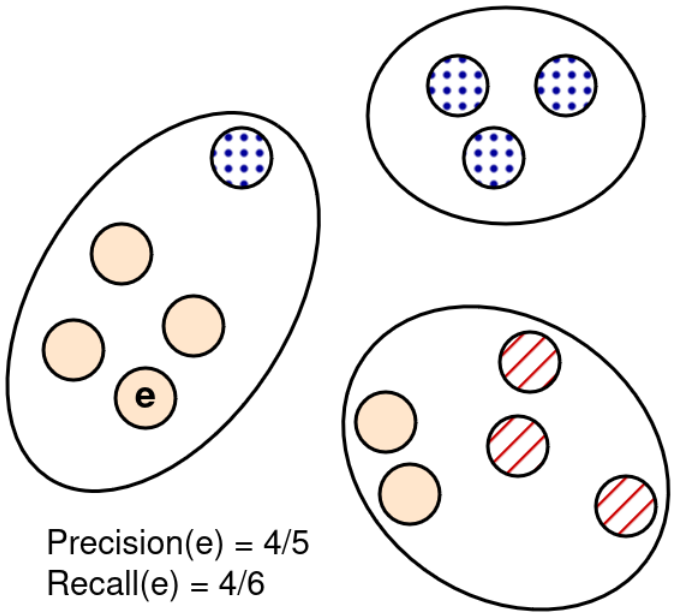


Figure C.1: Computation of BCubed Precision and Recall

The computation of BCubed Precision and Recall is illustrated in Figure C.1. Formally, for an entity e, let $L(e)$ denote the category (gold labels) of the entity and $C(e)$ be the assigned cluster. Correctness between a pair of entities checks that they are assigned the

same cluster if and only if they actually belong to the same category. In other words,

$$\text{Correctness}(e, e') = \begin{cases} 1 & C(e) = C(e') \iff L(e) = L(e') \\ 0 & \text{otherwise} \end{cases}$$

BCubed Precision of an item is the proportion of items in its cluster that have same category. Overall precision is the average of all items given as follows,

$$\text{BCubed Precision} = \text{Avg}_e \text{Avg}_{e':C(e')=C(e)} \text{Correctness}(e, e')$$

BCubed Recall is similarly defined,

$$\text{BCubed Recall} = \text{Avg}_e \text{Avg}_{e':L(e')=L(e)} \text{Correctness}(e, e')$$

BCubed F score follows the usual definition viz., the harmonic mean of BCubed Precision and BCubed Recall.