Gerhard Jäger*

Computational historical linguistics

https://doi.org/10.1515/tl-2019-0011

Abstract: Computational approaches to historical linguistics have been proposed for half a century. Within the last decade, this line of research has received a major boost, owing both to the transfer of ideas and software from computational biology and to the release of several large electronic data resources suitable for systematic comparative work.

In this article, some of the central research topics of this new wave of computational historical linguistics are introduced and discussed. These are *automatic assessment of genetic relatedness*, *automatic cognate detection*, *phylogenetic inference* and *ancestral state reconstruction*. They will be demonstrated by means of a case study of automatically reconstructing a Proto-Romance word list from lexical data of 50 modern Romance languages and dialects. The results illustrate both the strengths and the weaknesses of the current state of the art of automating the comparative method.

Keywords: historical linguistics, comparative method, phylogenetic inference

1 Introduction

Historical linguistics is the oldest sub-discipline of linguistics, and it constitutes an amazing success story. It gave us a clear idea of the laws governing language change, as well as detailed insights into the languages – and thus the cultures and living conditions – of prehistoric populations which left no written records. The diachronic dimension of languages is essential for a proper understanding of their synchronic properties. Also, the findings from historical linguistics are an important source of information for other fields of prehistory studies, such as archaeology, paleoanthropology and, in recent years, paleogenetics (Renfrew 1987; Pietrusewsky 2008; Anthony 2010; Haak et al. 2015, and many others).

The success of historical linguistics is owed to a large degree to a collection of very stringent methodological principles that go by the name of the *comparative method* (Meillet 1954; Weiss 2015). It can be summarized by the following workflow (from Ross and Durie 1996: 6–7):

***Corresponding author: Gerhard Jäger,** Institute of Linguistics, University of Tübingen, Wilhelmstraße 19, 72074 Tübingen, Germany, E-mail: gerhard.jaeger@uni-tuebingen.de

1. Determine on the strength of diagnostic evidence that a set of languages are genetically related, that is, that they constitute a 'family'.
2. Collect putative cognate sets for the family (both morphological paradigms and lexical items).
3. Work out the sound correspondences from the cognate sets, putting 'irregular' cognate sets on one side.
4. Reconstruct the protolanguage of the family as follows:
   a. Reconstruct the protophonology from the sound correspondences worked out in (3), using conventional wisdom regarding the directions of sound changes.
   b. Reconstruct protomorphemes (both morphological paradigms and lexical items) from the cognate sets collected in (2), using the protophonology reconstructed in (4a).
5. Establish innovations (phonological, lexical, semantic, morphological, morphosyntactic) shared by groups of languages within the family relative to the reconstructed protolanguage.
6. Tabulate the innovations established in (5) to arrive at an internal classification of the family, a 'family tree'.
7. Construct an etymological dictionary, tracing borrowings, semantic change and so forth, for the lexicon of the family (or of one language of the family).

In practice, it is not applied in a linear, pipeline-like fashion. Rather, the results of each intermediate step are subsequently used to inform earlier as well as later steps. This workflow is graphically depicted in Figure 1.

The steps (2)–(7) each involve a systematic, almost mechanical comparison and evaluation of many options such as cognacy relations, proto-form reconstructions or family trees. The first step, establishing genetic relatedness, is less regimented, but it generally involves a systematic comparison of many variables from multiple languages as well. It is therefore not surprising that there have been many efforts to formalize parts of this workflow to a degree sufficient to implement it on a computer.

Lexicostatistics (e.g. Swadesh 1952, 1955 and much subsequent work) can be seen as an early attempt to give an algorithmic rendering of step (6), even though it predates the computer age. Since the 1960s, several scholars applied computational methods within the overall framework of lexicostatistics (cf. e.g. Embleton 1986, *inter alia*). Likewise, there have been repeated efforts for computational treatments of other aspects of the comparative method, such as (Ringe 1992; Baxter and Manaster Ramer 2000; Kessler 2001) for step (1), (Kay 1964) for step (2), (Kondrak 2002) for steps (2) and (3), (Lowe and Mazaudon 1994) for steps (2) and

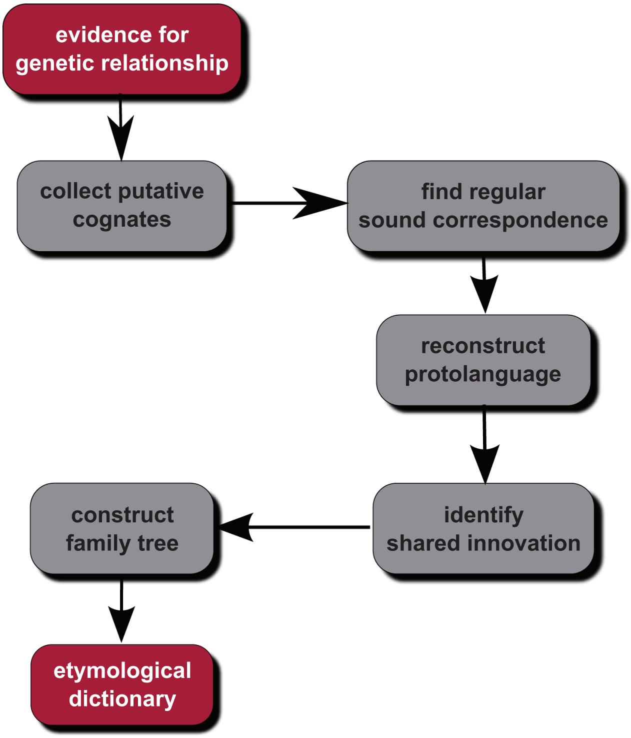
evidence for genetic relationship
collect putative cognates
find regular sound correspondence
reconstruct protolanguage
identify shared innovation
construct family tree
etymological dictionary

Figure 1: Workflow of the comparative method (according to Ross and Durie 1996).

(4), (Oakes 2000) for steps (2)–(7), (Covington 1996) for step (3), and (Ringe et al. 2002) for step (6), to mention just a few of the earlier contributions.

While the mentioned proposals mostly constitute isolated efforts of historical and computational linguists, the emerging field of computational historical linguistics received a major impetus since the early 2000s by the work of computational biologists such as Alexandre Bouchard-Côté, Russell Gray, Robert McMahon, Mark Pagel or Tandy Warnow and co-workers, who applied methods from their field to the problem of the reconstruction of language history, often in collaboration with linguists. This research trend might be dubbed *computational phylogenetic linguistics* as it heavily draws on techniques of *phylogenetic inference* from computational biology (Gray and Jordan 2000; Gray and Atkinson 2003; McMahon and McMahon 2005; Pagel et al. 2007; Atkinson et al. 2008; Gray et al. 2009; Dunn et al. 2011; Bouckaert et al. 2012; Bouchard-Côté et al. 2013; Pagel et al. 2013; Hruschka et al. 2015).

In recent years, more and more large collections of comparative linguistic data have become available in digital form, giving the field another boost. The following list gives a sample of the most commonly used databases; it is necessarily incomplete as new data sources are continuously made public.

- **Cognate-coded word lists**
  - *Indo-European Lexical Cognacy Database* (IELex; ielex.mpi.nl): collection of 225-concept Swadesh lists from 163 Indo-European languages (based on Dyen et al. 1992). Entries are given in orthography with manually assigned cognate classes; for part of the entries, IPA transcriptions are given.
  - *Austronesian Basic Vocabulary Database* (ABVD; Greenhill et al. 2008; language.psy.auckland.ac.nz/austronesian): collection of 210-item Swadesh lists for 1,467 languages from the Pacific region, mostly belonging to the Austronesian language family. Entries are given in phonetic transcription with manually assigned cognate classes.
- **Phonetically transcribed word lists**
  - *ASJP database* (compiled by the *Automatic Similarity Judgment Program*; Wichmann et al. 2016; asjp.clld.org): collection of word lists for 7,221 doculects (languages and dialects) over 40 concepts (100-item word lists for ca. 300 languages); entries are given in phonetic transcription.
- **Grammatical and typological classifications**
  - *World Atlas of Language Structure* (Haspelmath et al. 2008; wals.info): manual expert classifications of 2,679 languages and dialects according to 192 typological features.
  - *Syntactic Structures of the World's Languages* (sswl.railsplayground.net): Classification of 274 languages according to 148 syntactic features.
- **Expert language classifications**
  - *Ethnologue* (Lewis et al. 2016; https://www.ethnologue.com): genetic classification of 7,457 languages, alongside with information about number of speakers, location, and viability.
  - *Glottolog* (Hammarström et al. 2016; glottolog.org): genetic classification of 7,943 languages and dialects, alongside with information about geographic locations and extensive bibliographic references

Additionally there is a growing body of publicly available diachronic corpora of various languages, as well as studies using agent-based simulations of language change. The focus of this article is on computational work inspired by the comparative method, so this line of work will not further be covered here.

2 A program for *computational historical linguistics*

Conceived in a broad sense, computational historical linguistics comprises all efforts deploying computational methods to answer questions about the history of natural languages. As spelled out above, there is a decade-old tradition of this kind of research.

In this article, however, the term will be used in a rather narrower sense to describe an emerging subfield which has reached a certain degree of convergence regarding research goals, suitable data source and computational methods and tools to be deployed. I will use the abbreviation *CHL* to refer to computational historical linguistics in this narrow sense. The following remarks strive to describe this emerging consensus. They are partially programmatic in nature though; not all researchers active in this domain will agree with all of them.

CHL is informed by three intellectual traditions:

- the **comparative method** of classical historical linguistics,
- **computational biology**, especially regarding *sequence alignment* (cf. Durbin et al. 1989) and *phylogenetic inference* (see, e.g. Ewens and Grant 2005; Chen et al. 2014), and
- computational linguistics in general, especially modern statistical **natural language processing** (NLP).

CHL shares, to a large degree, the research objectives of the *comparative method*. The goal is to reconstruct the historical processes that led to the observed diversity of extant or documented ancient languages. This involves, *inter alia* establishing cognacy relations between words and morphemes, identifying regular sound correspondences, inferring family trees (*phylogenetic trees* or simply *phylogenies* in the biology-inspired terminology common in CHL), reconstructing proto-forms and historical processes such as sound laws and lexical innovations. CHL also shares some limitations of the comparative method, e.g. the idealized assumptions that language diversification proceeds in a tree-like fashion and that language contact is not systematically modeled.[1]

CHL's guiding model is adapted from *computational biology*, drawing on the observation going back to Schleicher and Darwin that the processes of biological and of linguistic diversification are analogous.[2] The history of a group of languages is represented by a phylogenetic tree (including branch lengths), with

[1] See, e.g. Heggarty et al. (2010); François (2015) for critical discussions of these limitations.

[2] See Atkinson and Gray (2005) for a detailed discussion of this analogy.

observed linguistic varieties at the leafs of the tree. Splits in a tree represent diversification events, i.e. the separation of an ancient language into daughter lineages. Language change is conceptualized as a continuous-time Markov process applying to discrete, finite-values characters. (Details will be spelled out below.) Inference amounts to finding the model (a phylogenetic tree plus a parameterization of the Markov process) that best explains the observed data.

Last but not least, CHL adopts techniques and methodological guidelines from *statistical NLP*. The pertinent computational tools, such as string comparison algorithms, to a certain degree overlap with those inspired by computational biology. Equally important are certain methodological standards from NLP and machine learning.

Generally, work in CHL is a kind of *inference*, where a collection of data is used as input (premises) to produce output data (conclusions). Input data can be phonetically or orthographically transcribed word lists, pairwise or multiply aligned word lists, grammatical feature vectors etc. Output data are for instance cognate class labels, alignments, phylogenies or proto-form reconstructions. Inference is performed by constructing a *model* and *training* its parameters. Following the standards in statistical NLP, the following guiding principles are desirable when performing inference:

- **Replicability.** All data used in a study, including all manual pre-processing steps, are available to the scientific community. Likewise, each computational inference step is either documented in sufficient detail to enable re-implementation, or made available as source code.
- **Rigorous evaluation.** The quality of the inference, or *goodness of fit* of the trained model, is evaluated by applying a well-defined quantitative measure to the output of the inference. This measure is applicable to competing models for the same inference task, facilitating model comparison and model selection.
- **Separation of training and test data.** Different data sets are used for training and evaluating a model.
- **Only raw data as input.** Only such data are used as input for inference that can be obtained without making prior assumptions about the inference task. For instance, word lists in orthographic or phonetic transcription are suitable as input if the transcriptions were produced without using diachronic information.

The final criterion is perhaps the most contentious one. It excludes, for instance, the use of orthographic information in languages such as English or French for training purposes, as the orthographic conventions of those languages reflect the phonetics of earlier stages. Also, it follows that the cognate class labels from databases such as IELex or ABVD, as well as expert classifications such as Ethnologue or Glottolog, are unsuitable as input for inference and should only be used as gold standard for training and testing.

Conceived this way, CHL is much narrower in scope than, e.g. computational phylogenetic linguistics. For instance, inference about the time depth and homeland of language families (such as Gray and Atkinson 2003; Bouckaert et al. 2012) is hard to fit into this framework as long as there are no independent test data to evaluate models against (but see Rama 2013). Also, it is common practice in computational phylogenetic linguistics to use manually collected cognate classifications as input for inference (Gray and Jordan 2000; Gray and Atkinson 2003; Pagel et al. 2007; Atkinson et al. 2008; Gray et al. 2009; Dunn et al. 2011; Bouckaert et al. 2012; Bouchard-Côté et al. 2013; Pagel et al. 2013; Hruschka et al. 2015). While the results obtained this way are highly valuable and insightful, they are not fully replicable, since expert cognacy judgments are necessarily subjective and variable. Also, the methods used in the work mentioned do not generalize easily to understudied language families, since correctly identifying cognates between distantly related languages requires the prior application of the classical comparative method, and the necessary research has not been done with equal intensity for all language families.

3 A case study: reconstructing Proto-Romance

In this section, a case study will be presented that illustrates many of the techniques common in current CHL. Training data are 40-item word lists from 50 Romance (excluding Latin) and 3 Albanian[3] languages and dialects in phonetic transcription from the ASJP database (Wichmann et al. 2016) (version 17, accessed on August 2, 2016 from asjp.clld.org/static/download/asjp-dataset.tab.zip). The inference goal is the reconstruction of the corresponding word list from the latest common ancestor of the Romance languages and dialects (Proto-Romance, i.e. some version of Vulgar Latin). The results will be tested against the Latin word lists from ASJP.[4] A subset of the data used is shown in Table 1 for illustration. The phonetic transcriptions use the 41 ASJP sound classes (cf. Brown et al. 2013).

[3] The inclusion of Albanian will be motivated below.

[4] As discussed below, the latest common ancestor of the modern Romance varieties is Vulgar Latin rather than classical Latin. I use the latter here for evaluation purposes because it is the closest proxy for Vulgar Latin that is contained in the data base.

Table 1: Sample of word lists used.

Concept	ALBANIAN	SPANISH	ITALIAN	ROMANIAN	LATIN
horn	bri	kerno	korno	korn	kornu
knee	Tu	rodiya	jinokkyo	jenuNk	genu
mountain	mal	sero	monta5a	munte	mons
liver	m3lCi	igado	fegato	fikat	yekur
we	ne	nosotros	noi	noi	nos
you	ju	ustet	tu	tu	tu
person	vet3	persona	persona	persoan3	persona
louse	morr	pioho	pidokko	p3duke	pedikulus
new	iri	nuevo	nwovo	nou	nowus
hear	d3gyoy	oir	ud	auz	audire
sun	dyell	sol	sole	soare	sol
tree	dru	arbol	albero	pom	arbor
breast	kraharor	peCo	pEtto	pept	pektus
drink	pirye	bebe	bere	bea	bibere
hand	dor3	mano	mano	m3n3	manus
die	vdes	mori	mor	mur	mori
name	em3r	nombre	nome	nume	nomen
eye	si	oho	okkyo	ok	okulus

Diacritics are removed. If the database lists more than one translation for a concept in a given language, only the first one is used.[5]

The following steps will be performed (mirroring to a large degree the steps of the comparative method):

1. Demonstrate that the Romance languages and dialects are related.
2. Compute pairwise string alignments and string similarities between synonymous words from different languages/dialects.
3. Cluster the words for each concept into automatically inferred cognate classes.
4. Infer a phylogenetic tree (or a collection of trees).
5. Perform ancestral state reconstruction for cognate classes to infer the cognate class of the Proto-Romance word for each concept.
6. Perform multiple sequence alignment of the reflexes of those cognate classes within the Romance languages and dialects.
7. Perform ancestral state reconstruction to infer the state (sound class or gap) of each column in the multiple sequence alignments.
8. Compare the results to the Latin ASJP word list.

[5] This choice was made to keep the algorithmic treatment simple. A fuller model would factor in the role of synonymy.

3.1 Demonstration of genetic relationship

In (Jäger 2013), a dissimilarity measure between ASJP word lists is developed. Space does not permit to explain it in any detail here. Suffice it to say that this measure is based on the average string similarity between the corresponding elements of two word lists while controlling for the possibility of chance similarities. Let us call this dissimilarity measure between two word lists the *PMI distance*, since it makes crucial use of the *pointwise mutual information* (PMI) between phonetic strings.

To demonstrate that all Romance languages and dialects used in this study are mutually related, I will use the ASJP word lists from Papunesia, i.e. "all islands between Sumatra and the Americas, excluding islands off Australia and excluding Japan and islands to the North of it" (Hammarström et al. 2016) as training data and the ASJP word lists from Africa as test data.[6] Input for inference are PMI distances between pairs of languages/dialect, and the output is the classification of this pair as *related* or *unrelated*, where two doculects count as related if they belong to the same language family according to the Glottolog classification. The distribution of PMI distances are shown in Figure 2. The graphics illustrates that all doculect pairs with a PMI distance $\leq$ 0.75 are, with a very high probability, related. The largest PMI distance among Romance dialects (between Aromanian and Nones) is 0.65.

A statistical test confirms this impression. I fitted a cumulative density estimation for the PMI distances of the unrelated doculect pairs from the training data, using the R-package *logspline* (Kooperberg 2016). If a pair of doculects has a PMI distance d, the value of the cumulative density function for d can then be interpreted as the (one-sided) p-values for the null hypothesis that the doculects are unrelated.

Using a threshold of $\alpha = 0.0001$, I say that a doculect pair is predicted to be related if the model predicts it to be unrelated with a probability $\leq \alpha$. In Table 2, the predictions are tabulated against the Glottolog gold standard. These results amount to ca. 0.3% of false positives and ca. 84% of false negatives. This test ensures that the chosen model and threshold is sufficiently conservative to keep the risk of wrongly assessing doculects to be related small. Since the method is so conservative, it produces a large amount of false negatives.

In the next step, I compute the probability of all pairs of Romance doculects to be unrelated, using the model obtained from the training data. Using the Holm–Bonferroni method to control for multiple tests, the highest p-value for the null

[6] I chose different macro-areas for training and testing to minimize the risk that the data are non-independent due to common ancestry or language contact.

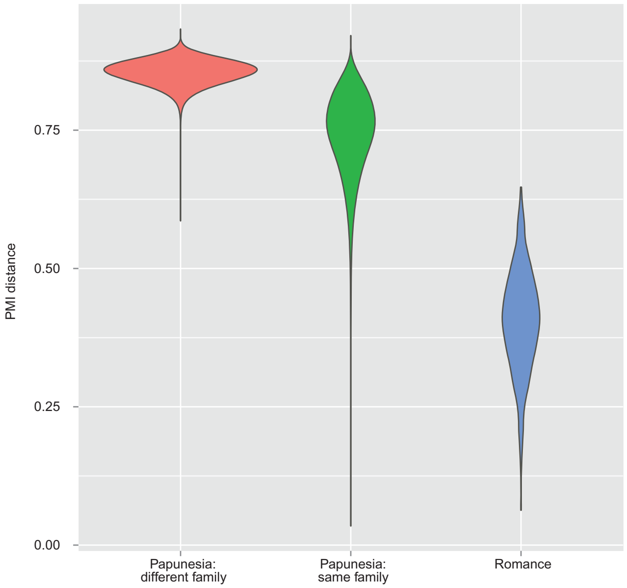

Figure 2: PMI distances between related and unrelated doculects from Papunesia, and between the Romance doculects.

Table 2: Contingency table of gold standard versus prediction for the test data of doculect pairs.

	Glottolog:	
	Unrelated	**Related**
Prediction: unrelated	1,254,726	787,023
Prediction: related	532	153,279

hypothesis that the doculect pair in question is unrelated is 1.8×10^{-5}, i.e. all adjusted p-values are $< \alpha$. We can therefore reject the null hypothesis for all Romance doculect pairs.

3.2 Pairwise string comparison

All subsequent steps rely on a systematic comparison of words for the same concept from different doculects. Let us consider as an example the words for *water* from Catalan and Italian from ASJP, “aigua” and “acqua”. Both are descendants of the Latin “aqua” (Meyer-Lübke 1935: 46). In ASJP transcription, these are `aixw~3` and `akwa`. The sequence `w~` in the Catalan word encodes a diacritic (indicating labialization of the preceding segment) and is removed in the subsequent processing steps.

A *pairwise sequence alignment* of two strings arranges them in such a way that corresponding segments are aligned, possibly inserting gap symbols for segments in one string that have no correspondent in the other string. For the example, the historically correct alignment would arguably be as follows:

```
aix-3
a-kwa
```

In this study, the quality of a pairwise alignment is quantified as its aggregate *pointwise mutual information* (PMI). (See List 2014 for a different approach.) The PMI between two sound classes a, b is defined as

$$PMI(a,b) \doteq \log \frac{s(a,b)}{q(a)q(b)}. \tag{1}$$

Here $s(a,b)$ is the probability that a is aligned to b in a correct alignment, and $q(a), q(b)$ are the probabilities of occurrence of a and b in a string. If one of the two symbols is a gap, the PMI score is a *gap penalty*. I use *affine gap penalties*, i.e. the gap penalty is reduced if the gap is preceded by another gap. The captures the heuristics that both addition and deletion of phonetic material often targets several consecutive segments.

If the PMI scores for each pair of sound classes and the gap penalties are known, the best alignment between two strings (i.e. the alignment maximizing the aggregate PMI score) can efficiently be computed using the Needleman–Wunsch algorithm (Needleman and Wunsch 1970).

The quantities $s(a,b)$ and $q(a), q(b)$ must be estimated from the data. Here I follow a simplified version of the parameter estimation technique from (Jäger 2013). In a first step, I set

$$PMI_0(a,b) \doteq \begin{cases} 0 \text{ if } a = b \\ -1 \text{ else.} \end{cases}$$

Also, I set the initial gap penalties to −1. (This amounts to *Levenshtein alignment*.) For instance, the alignment for the example given above would come out as

```
aix3
akwa
```

This alignment has one identity (the initial `a`) and three non-identities, so the PMI_0 score is −3.

Using these parameters, all pairs of words for the same concept from different doculects are aligned.

From those alignments, $s(a, b)$ is estimated as the relative frequency of a and b being aligned among all non-gap alignment pairs, while $q(a)$ is estimated as the relative frequency of sound class a in the data. The PMI scores are then estimated using equation (1). For the gap penalties I used the values from (Jäger 2013), i.e. −2.49 for opening gaps and −1.70 for extending gaps (i.e. for gaps preceded by another gap). Using those parameters, all synonymous word pairs are realigned.

In the next step, only word pairs with an aggregate PMI score ≥ 4.45 are used. (This threshold is taken from Jäger 2013 as well.) Those word pairs are realigned and the PMI scores are reestimated. This step is repeated ten times.

The threshold of 4.45 is rather strict; almost all word pairs above this threshold are either cognates or loans. For instance, for the language pair Italian/Albanian, the only translation pair with a higher PMI score is Italian `peSe`/Albanian `peSk` ("fish"), where the former is a descendant and the latter a loan from Latin *piscis* (cf. http://ielex.mpi.nl). For Spanish/Romanian, two rather divergent Romance languages, we find eight such word pairs. They are shown alongside with the inferred alignments in Table 3.

The aggregate PMI score for the best alignment between two strings is a measure for the degree of similarity between the strings. I will call it the *PMI similarity* henceforth.

Table 3: Word pair alignments from Spanish and Romanian.

Concept	Alignment	PMI score
person	`perso-na` `persoan3`	14.23
tooth	`diente` `di-nte`	10.13
blood	`sangre` `s3nj-e`	8.04
hand	`mano` `m3n3`	6.71
one	`uno` `unu`	5.61
die	`mori` `mur-`	5.16
come	`veni` `ven-`	5.01
name	`nombre` `num-e`	4.98

3.3 Cognate clustering

Automatic cognate detection is an area of active investigation in CHL (Dolgopolsky 1986; Bergsma and Kondrak 2007; Hall and Klein 2010; Turchin et al. 2010; Hauer and Kondrak 2011; List 2012, 2014; Rama 2015; Jäger and Sofroniev 2016; Jäger et al. 2017, *inter alia*). For the present study, I chose a rather simple approach based on unsupervised learning.

Figure 3 shows the PMI similarities for words from different doculects having different or identical meanings. Within our data, synonymous word pairs are, on average, more similar to each other than non-synonymous ones. The most plausible explanation for this effect is that the synonymous word pairs contain a large proportion of cognate pairs. Therefore "identity of meaning" will be used as a proxy for "being cognate". The implicit assumption underlying this procedure is that cognate words always have the same meaning. This is evidently false when considering the entire lexicon. There is a plethora of examples, such as English *deer* vs. German *Tier* "animal", which are cognate (cf. Kroonen 2013: 94) without being synonyms. However, within the 40-concept core vocabulary space covered by ASJP, such cross-concept cognate pairs are arguably very rare.

I fitted a logistic regression with PMI similarity as independent and synonymy as dependent variables.

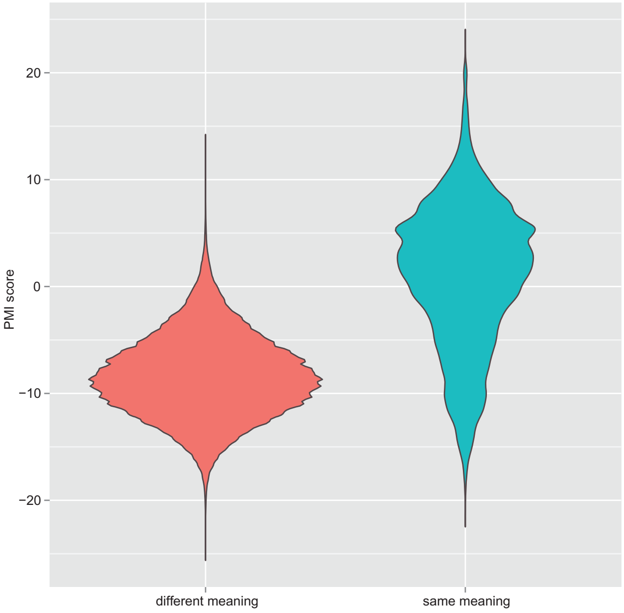

Figure 3: PMI similarities for synonymous and non-synonymous word pairs.

For each concept, a weighted graph is constructed, with the words denoting this concept as vertices. Two vertices are connected if the predicted probability of these words to be synonymous (based on their PMI similarity and the logistic regression model) is ≥ 0.5. The weight of each edge equals the predicted probabilities. The nodes of the graph are clustered using the weighted version of the *Label Propagation* algorithm (Raghavan et al. 2007) as implemented in the *igraph* software (Csardi and Nepusz 2006). As a result, a *class label* is assigned to each word. Non-synonymous words never carry the same class label. Table 4 illustrates

Table 4: Automatic cognate clustering for concept "person".

Doculect	Word	Class label
ALBANIAN	vet3	0
ALBANIAN_TOSK	vEt3	0
ARAGONESE	ombre	1
ITALIAN_GROSSETO_TUSCAN	omo	2
ROMANIAN_MEGLENO	wom	2
VLACH	omu	2
ASTURIAN	persona	3
BALEAR_CATALAN	p3rson3	3
CATALAN	p3rson3	3
FRIULIAN	pErsoN	3
ITALIAN	persona	3
SPANISH	persona	3
VALENCIAN	persone	3
CORSICAN	nimu	4
DALMATIAN	om	5
EMILIANO_CARPIGIANO	om	5
ROMANIAN_2	om	5
TURIA_AROMANIAN	om	5
EMILIANO_FERRARESE	styan	6
LIGURIAN_STELLA	kristyaN	6
NEAPOLITAN_CALABRESE	kr3styan3	6
ROMAGNOL_RAVENNATE	sCan	6
ROMANSH_GRISHUN	k3rSTawn	6
ROMANSH_SURMIRAN	k3rstaN	6
GALICIAN	ome	7
GASCON	omi	7
PIEMONTESE_VERCELLESE	omaN	8
ROMANSH_VALLADER	uman	8
ALBANIAN_GHEG	5eri	9
SARDINIAN_CAMPIDANESE	omini	9
SARDINIAN_LOGUDARESE	omine	9

the resulting clustering for the concept "person" and a subset of the doculects. A manual inspection reveals that the automatic classification does not completely coincide with the cognate classification a human expert would assume. For instance, the descendants of Latin *homo* are split into classes 1, 2, 5, and 7. Also, Gheg Albanian `5eri` and Sardinian `omini` have the same label but are not cognate.

Based on evaluations against manually assembled cognacy judgments for different but similar data (Jäger and Sofroniev 2016; Jäger et al. 2017), we can expect an average F-score of 60%–80% for automatic cognate detection. This means that on average, for each word, 60%–80% of its true cognates are assigned the same label, and 60%–80% of the words carrying the same label are genuine cognates.

3.4 Phylogenetic inference

3.4.1 General remarks

A *phylogenetic tree* (or simply *phylogeny*) is a data structure similar to family trees according to the comparative method, but there are some subtle but important differences between those concepts. Like a family tree, a phylogeny is a tree graph, i.e. an acyclic graph. If one node in the graph is identified as *root*, the phylogeny is *rooted*; otherwise it is *unrooted*. The branches (or edges) of a phylogeny have non-negative *branch lengths*. A phylogeny without branch length is called *topology*.

Nodes with a degree 1 (i.e. nodes which are the endpoint of exactly one branch) are called *leaves* or *tips*. They are usually labeled with the names of observed entities, such as documented languages. Nodes with a degree > 1 are the *internal nodes*. If the root (if present) has degree 2 and all other internal nodes have degree 3 (i.e. one mother node and two daughter nodes), the phylogeny is *binary-branching*. Most algorithms for phylogenetic inference produce binary-branching trees.

Like a linguistic family tree, a rooted phylogeny is a model of the historic process leading to the observed diversity between the objects at the leaves. Time flows from the root to the leaves. Internal nodes represent unobserved historical objects, such as ancient languages. Branching nodes represent *diversification events*, i.e. the splitting of a lineage into several daughter lineages.

The most important difference between family trees and phylogenies is the fact that the latter have branch lengths. Depending on the context, these lengths may represent two different quantities. They may capture the historic time (measured in years) between diversification events, or they indicate the *amount of change* along the branch, measured for instance as the expected number of lexical replacements or the expected number of sound changes. The two interpretations only coincide if the rate of change is constant. This assumption is known to be invalid for language change (cf. e.g. the discussion in McMahon and McMahon 2006).

Another major difference, at least in practice, between family trees and phylogenies concerns the type of justification that is expected for the stipulation of an internal node. According to the comparative method, such a node is justified if and only if a *shared innovation* can be reconstructed for all daughter lineages of this node.[7] Consequently, family trees obtained via the comparative method often contain multiply branching nodes because the required evidence for further subgrouping is not available. Phylogenies, in contradistinction, are mostly binary branching, at least in practice. Partially this is a matter of computational convenience. Since the set of binary branching trees is a proper subset of the set of all trees, this reduces the search space. Also, algorithms working recursively leaves-to-root can be formulated in a more efficient way if all internal nodes are known to have at most two daughters. Furthermore, the degree of justification of a certain topology is evaluated *globally*, not for each internal node individually. In the context of phylogenetic inference, it is therefore not required to identify shared innovations for individual nodes.

There is a large variety of algorithms from computational biology to infer phylogenies from observed data. The overarching theme of *phylogenetic inference* is that a phylogeny represents (or is part of) a mathematical model explaining the observed variety. There are criteria quantifying how good an explanation a phylogeny provides for observed data. Generally speaking, the goal is to find a phylogeny that provides an optimal explanation for the observed data. The most commonly used algorithms are (in ascending order of sophistication and computational costs) *Neighbor Joining* (Saitou and Nei 1987) and its variant *BIONJ* (Gascuel 1997), *FastMe* (Desper and Gascuel 2002), *Fitch-Margoliash* (Fitch and Margoliash 1967), *Maximum Parsimony* (Fitch 1971), *Maximum Likelihood*[8] and *Bayesian Phylogenetic Inference* (cf. Chen et al. 2014 for an overview).

The latter two approaches, *Maximum Likelihood* and *Bayesian Phylogenetic Inference*, are based on a probabilistic model of language change. To apply them, a language has to be represented as a *character vector*. A *character* is a feature with a finite number of possible values, such as “order of verb and object”, “the first person plural pronoun contains a dental consonant” or what have you. In

[7] “The only generally accepted criterion for subgrouping is *shared innovation*.” (Campbell 2013: 175, emphasis in original).

[8] This method was developed incrementally; Edwards and Cavalli-Sforza (1964) is an early reference.

most applications, characters are binary, with "0" and "1" as possible values. In what follows, we will assume all characters are binary.

Diachronic change of a character value is modeled as a *continuous time Markov process*. At each point in time, a character can spontaneously switch to the other value with a fixed probability density. A two-state process is characterized by two parameters, r and s, where r is the *rate of change* of 0 → 1 (the probability density of a switch to 1 if the current state is 0) and s the rate of change for 1 → 0. For a given time interval of length t, the probability of being in state i at the start of the interval and in state j at the end is then given by $P(t)_{ij}$, where

$$P(t) = \frac{1}{r+s}\begin{pmatrix} s + re^{-(r+s)t} & r - re^{-(r+s)t} \\ s - se^{-(r+s)t} & r + se^{-(r+s)t} \end{pmatrix}.$$

For instance, the cell in the upper right corner of the matrix, multiplied with the coefficient $1/(r+s)$, gives the probability that the system is in state 1 after a time interval of length t if it is in state 0 at the beginning of the interval. This probability is a function of t that starts at 0 for $t = 0$ and exponentially converges towards $r/(r+s)$ as t gets longer. The cell in the upper left corner gives a probability that the system is still or again in state 0 after time t if it starts in state 0. (The possibility of multiple switches occurring during the interval is factored in.) For $t = 0$, this probability is 1, and it exponentially converges to $s/(r+s)$ as t grows to infinity. The same holds, *mutatis mutandis*, for the second row.

If the state at the end of the interval t is known – for instance by observation –, one can try to probabilistically estimate the state at the beginning of the interval. However, the longer t is, the less precise will this estimate be. If t grows to infinity, the amount of information that can be inferred about the initial state converges to 0.

A probabilistic model for a given set of character vectors is a phylogenetic tree (with the leaves indexed by the characters vectors) plus a mapping from edges to rates (r, s) for each character and a probability distribution over character values at the root for each character.

Suppose we know not only the character states at the leaves of the phylogeny but also at all internal nodes. The likelihood of a given branch is then given by $P(t)_{ij}$, where i and j are the states at the top and the bottom of the branch respectively, and t is the length of the branch. The likelihood of the entire phylogeny for a given character is then the product of all branch likelihoods, multiplied with the probability of the root state. The total likelihood of the phylogeny is the product of its likelihoods for all characters.

If only the character values for the leaves are known, the likelihood of the phylogeny given the character vectors at the leaves is the sum of its likelihoods for all possible state combinations at the internal nodes.

This general model is very parameter rich since for each branch and each character, a pair of rates has to be specified. There are various ways to reduce the degrees of freedom. The simplest method is to assume that rates are constant across branches and characters, and that the root probability of each character equals the equilibrium probabilities of the Markov process: $({}^{s}/_{(r+s)}, {}^{r}/_{(r+s)})$. More sophisticated approaches assume that rates vary across characters and across branches according to some parameter-poor probability distribution, and the expected likelihood of the tree is obtained by integrating over this distribution. For a detailed mathematical exposition, the interested reader is referred to the relevant literature from computational biology, such as Ewens and Grant (2005).

A parameterized model, i.e. a phylogeny plus rate specifications for all characters and branches, and root probabilities for each character, assigns a certain likelihood to the observed character vectors. *Maximum Likelihood* (ML) inference searches for the model that maximizes this likelihood given the observations. While the optimal numerical parameters of a model, i.e. branch lengths, rates and root probabilities, can efficiently be found by standard optimization techniques, finding the topology that gives rise to the ML-model is computationally hard. Existing implementations use various heuristics to search the tree space and find some local optimum, but there is no guarantee that the globally optimal topology is found.[9]

Bayesian phylogenetic inference requires some suitable prior probability distributions over models (i.e. topologies, branch lengths, rates, possibly rate variations across characters and rate variation across branches) and produces a sample of the posterior distribution over models via a Markov Chain Monte Carlo simulation.[10]

3.4.2 Application to the case study

For the case study, doculects were represented by two types of binary characters:

- **Inferred class label characters** (cf. Subsection 3.3). Each inferred class label is a character. A doculect has value 1 for such a character if and only if its word list contains a word carrying this label.[11]

[9] Among the best software packages currently available for ML phylogenetic inference are *RAxML* (Stamatakis 2014) and *IQ-Tree* (Nguyen et al. 2015).

[10] Suitable software packages are, *inter alia*, *MrBayes* (Ronquist and Huelsenbeck 2003) and *BEAST* (Bouckaert et al. 2014).

[11] If a word list contains no entry for a certain concept, all characters pertaining to this concept are undefined for this concept. The same principle applies to the soundclass-concept characters. Leaves with undefined character values are disregarded when computing the likelihood of a phylogeny for that character.

- **Soundclass-concept characters**. There is one character for each pair (s, c) of a sound class s and a concept c. A doculect has value 1 for that character if and only if its word list contains a word w for c that contains s in its transcription.

Both types of characters carry a diachronic signal. For instance, the mutation 0 → 1 for class label 6/concept *person* (cf. Table 4) represents a lexical replacement of Latin "homo" or "persona" by descendants of Latin "christianus" in some Romance dialects (Meyer-Lübke 1935: 179). The mutation 0 → 1 for the soundclass-concept character k/*person* represents the same historical process. Soundclass-concept characters, however, also capture sound shifts. For instance, the mutation 0 → 1 for b/*person* reflects the epenthetic insertion of b in descendants of Latin "homo" in some Iberian dialects.[12]

I performed Bayesian phylogenetic inference on those characters. The inference was carried out using the Software *MrBayes* (Ronquist and Huelsenbeck 2003). Separate rate models were inferred for the two character types. Rate variation across characters was modeled by a discretized Gamma distribution using four rate categories. I assumed no rate variation across edges. Root probabilities were identified with equilibrium probabilities. An ascertainment correction for missing all-0 characters was used.

I assumed rates to be constant across branches. This entails that the fitted branch lengths reflect the expected amount of change (i.e. the expected number of mutations) along that branch.

In such a model, the likelihood of a phylogeny does not depend on the location of the root (the assumed Markov process is *time reversible*.) Therefore, phylogenetic inference provides no information about the location of the root. This motivates the inclusion of the Albanian doculects. Those doculects were used as *outgroup*, i.e. the root was placed on the branch separating the Albanian and the Romance doculects.

I obtained a sample of the posterior distribution containing 2,000 phylogenies. Figure 4 displays a representative member of this sample (the *maximum clade credibility* tree). The labels at the nodes indicate *posterior probabilities* of that node, i.e. the proportion of the phylogenies in the posterior sample having the same sub-group.

These posterior probabilities are mostly rather low, indicating a large degree of topological variation in the posterior sample. Some subgroups, such as Balkan Romance or the Piemontese dialects, achieve high posterior probabilities though.

[12] It should be pointed out that the method used here only detects sound changes in individual lexical items, not regular sound shifts in the Neogrammarian's sense.

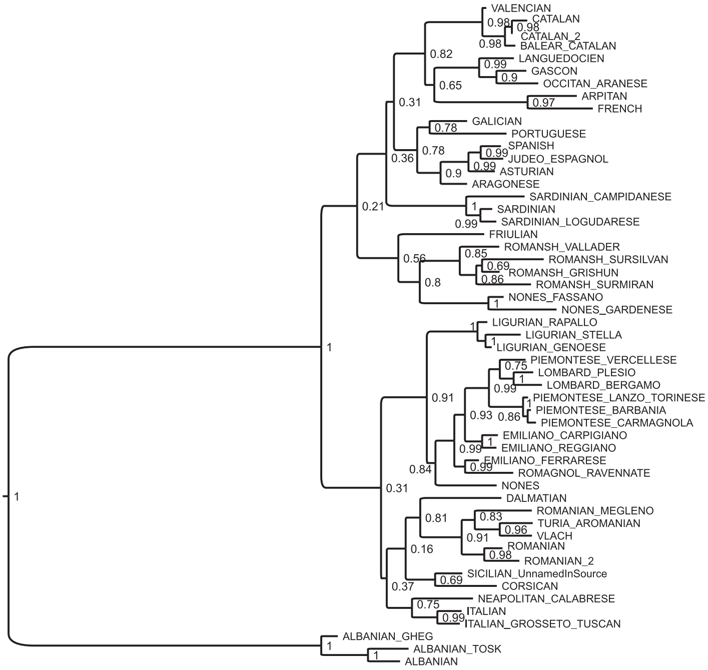

Figure 4: Representative phylogeny from the posterior distribution. Labels at the internal nodes indicate posterior probabilities.

Notably, branch lengths carry information about the amount of change. According to the phylogeny in Figure 4, for instance, the Tuscan dialect of Italian (ITALIAN_GROSSETO_TUSCAN) is predicted to be the most conservative Romance dialect (since its distance to the latest common ancestors of all Romance dialects is shortest) and French the most innovative one.

These results indicate that the data only contain a weak tree-like signal. This is unsurprising since the Romance languages and dialects form a dialect continuum where horizontal transfer of innovations is an important factor.

Phylogenetic trees, like traditional family trees, only model vertical descent, not horizontal diffusion. They are therefore only an approximation of the historical truth. But nevertheless, they are useful as statistical models for further inference steps.

3.5 Ancestral state reconstruction

If a fully specified model is given, it is possible to estimate the probability distributions over character states for each internal node.

Let $M = \langle \mathcal{T}, \vec{\theta} \rangle$ be a model, i.e. a phylogeny $\mathcal{T}$ plus further parameters $\vec{\theta}$ (rates and root probabilities, possibly specifications of rate variation). Let i be a character and n a node within $\mathcal{T}$.

The parameters $\vec{\theta}$ specify a Markov process, including rates, for the branch leading to n. Let $\langle \pi_0, \pi_1 \rangle$ be the equilibrium probabilities of that process. (If n is the root, $\langle \pi_0, \pi_1 \rangle$ are directly given by $\vec{\theta}$.)

Let $M(n_i = x)$ be the same model as M, except that the value of character i at node n is fixed to the value x. $\mathcal{L}(M)$ is the likelihood of model M given the observed character vectors for the leaves.

The probability distribution over values of character i at node n, given M, is determined by Bayes' rule:

$$P(n_i = x|M) = \frac{\mathcal{L}(M(n_i = x)) \times \pi_x}{\sum_y (\mathcal{L}(M(n_i = y)) \times \pi_y)}$$

Figure 5 illustrates this principle with the Romance part of the tree from Figure 4 and the character *person*:3 (cf. Table 4). The pie charts at the nodes display the probability distribution for that node, where white represents 0 and red 1.

This kind of computation was carried out for each class label character and each tree in the posterior sample for the latest common ancestor of the Romance doculects. For each concept, the class label for that concept with the highest average probability for value 1 at the root of the Romance subtree was inferred to represent the cognate class of the Proto-Romance word for that concept.[13] For the concept *person*, e.g. character *person*:3 (representing the descendants of Latin "persona"), comes out as the best reconstruction.

3.6 Multiple sequence alignment

In the previous step, for the concept *eye*, the class label 6 was reconstructed for Proto-Romance. Its reflexes are given in Table 5.

A *multiple sequence alignment* (MSA) is a generalization of pairwise alignment to more than two sequences. Ideally, all segments within a column are descendants of the same sound in some common ancestor.

MSA, as applied to DNA or protein sequences, is a major research topic in bioinformatics. The techniques developed in this field are *mutatis mutandis* also applicable to MSA of phonetic strings. In this subsection, one approach will briefly be sketched. For a wider discussion and proposals for related but different approaches, see (List 2014).

[13] See (Jäger and List 2017) for further elaboration and justification of this method of ancestral state reconstruction.

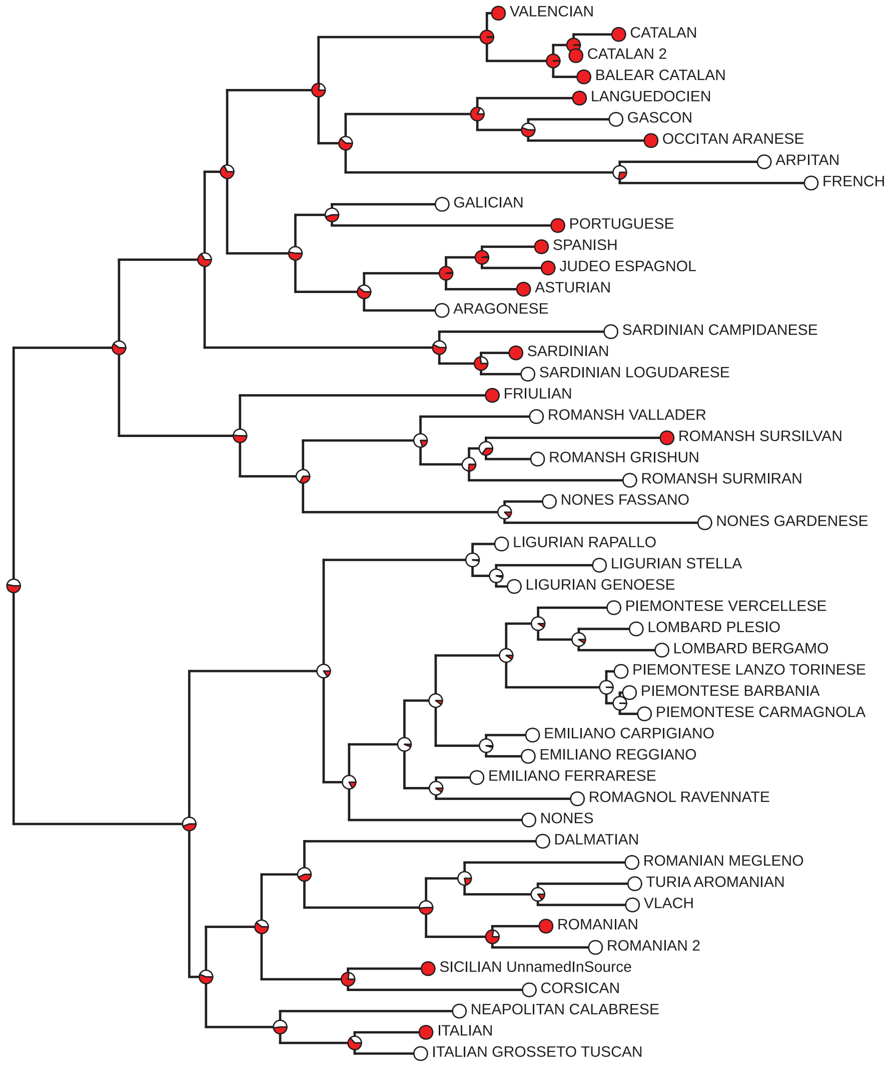

Figure 5: Ancestral state reconstruction for character *person*:3.

Here I will follow the overall approach from (Notredame et al. 2000) and combine it with the PMI-based method for pairwise alignment described in Subsection 3.2. (Notredame et al. 2000) dub their approach *T-Coffee* ("Tree-based Consistency Objective Function For alignment Evaluation"), and I will use this name for the method sketched here as well.

Table 5: Reflexes of class label *eye*:6.

Doculect	Word
DALMATIAN	vaklo
ITALIAN	okkyo
ITALIAN_GROSSETO_TUSCAN	okyo
NEAPOLITAN_CALABRESE	wokyo
ROMANIAN_2	oky
ROMANIAN_MEGLENO	wokLu
SARDINIAN_LOGUDARESE	okru
SICILIAN_UnnamedInSource	okiu
TURIA_AROMANIAN	okLu
VLACH	okklu

In a first step, all pairwise alignments between words from the list to be multiply aligned are collected. For this purpose, I use PMI pairwise alignment. Some examples would be

okiu	vaklo	okkyo	-okyo	o-ky-	okru
oky-	wokLu	o-ky-	wokyo	okklu	okiu
0.67	0.2	1.0	1.0	0.67	0.75

The last row shows the *score* of the alignment, i.e. the proportion of identical matches (disregarding gaps).

In a second step, all *indirect alignments* between a given word pair are collected, which are obtained via relation composition with a third word. Some examples for indirect alignments between okiu and oky would be:

okiu	-okiu	okiu	-okiu	oki-u
okyo	wokyo	oky-	wokLu	okklu
oky-	-oky-	oky-	-oky-	o-ky-

The direct pairwise alignment matches the *i* in okiu with the y in oky. Most indirect alignments pair these two positions as well, but not all of them. In the last columns, the i from okiu is related to the k of oky, and the y from oky with a gap. For each pair of positions in two strings, the relative frequency of them being indirectly aligned, weighted by the score of the two pairwise alignments relating them, is summed. They form the *extended score* between those positions.

The optimal MSA for the entire group of words is the one where the sum of the pairwise extended scores per column is maximized. Finding this global optimum is computationally not feasible though, since the complexity of this task grows exponentially with the number of sequences. *Progressive alignment* (Hogeweg and Hesper 1984) is a method to obtain possibly sub-optimal but good MSAs in polynomial time. Using a *guide tree* with sequences at the leaves, MSAs are obtained recursively leaves-to-root. For each internal node, the MSAs at the daughter nodes are combined via the Needleman–Wunsch algorithm while respecting all partial alignments from the daughter nodes.

For the words from Table 5, this method produces the MSA in Table 6. The tree in Figure 4, pruned to the doculects represented in the word lists, was used as guide tree.

Using this method, MSAs were computed for each inferred class label that was inferred to be present in Proto-Romance via Ancestral State Reconstruction.

3.7 Proto-form reconstruction

A final step toward the reconstruction of Proto-Romance forms, *Ancestral State Reconstruction* is performed for the sound classes in each column, for each MSA obtained in the previous step.

Consider the first column of the MSA in Table 6. It contains three possible states, v, w, and the gap symbols -. For each of these states, a binary presence–absence character is constructed. For doculects which do not occur in the MSA in question, this character is undefined.

The method for ancestral state reconstruction described in Subsection 3.5 was applied to these characters. For phylogeny in the posterior sample, the probabilities for state 1 at the Proto-Romance node was computed for each character. For each column of an MSA, the state with the highest average probability was considered as reconstructed.

Table 6: Multiple Sequence Alignment for the word from Table 5, using the tree from Figure 4 as guide tree.

Doculect	**Alignment**
DALMATIAN	va-klo
ITALIAN	-okkyo
ITALIAN_GROSSETO_TUSCAN	-o-kyo
NEAPOLITAN_CALABRESE	wo-kyo
ROMANIAN_2	-o-ky-
ROMANIAN_MEGLENO	wo-kLu
SARDINIAN_LOGUDARESE	-o-kru
SICILIAN_UnnamedInSource	-o-kiu
TURIA_AROMANIAN	-o-kLu
VLACH	-okklu

The reconstructed proto-form for a given concept is then obtained by concatenating the reconstructed states for the corresponding MSA and deleting all gap symbols. The results are given in Table 7.

3.8 Evaluation

To evaluate the quality of the automatic reconstructions, they were compared to the corresponding elements of the Latin word list. For each reconstructed word, the normalized Levenshtein distance (i.e. the Levenshtein distance divided by the length of the longer string) to each Latin word (without diacritics) for that concept was computed. The smallest such value counts as the score for that concept. The average score was 0.484. The extant Romance doculects have an average score of 0.627. The most conservative doculect, Sardinian, has a score of 0.502, and the least conservative, Arpitan, 0.742. The evaluation results are depicted in Figure 6.

These findings indicate that the automatic reconstruction does in fact capture a historical signal. Manual inspection of the reconstructed word list reveals that, to a large degree, the discrepancies to Latin actually reflect language change between Classical Latin and the latest common ancestor of the modern Romance doculects, namely Vulgar Latin.[14] To mention just a few points: (1) Modern Romance nouns are mostly derived from the (Vulgar-)Latin accusative form (Herman 2000: 3), while the word lists contains the nominative form. For instance, the common ancestor forms for “tooth” and “night” are *dentem* and *noctem*. The reconstructed `t` in the corresponding reconstructed forms are therefore historically correct. (2) Some Vulgar Latin words are morphologically derived from their Classical Latin counterparts, such as *mons* → *montanea* “mountain” (Meyer-Lübke 1935: 464) or *genus* → *genukulum* “knee” (Meyer-Lübke 1935: 319). This is likewise partially reflected in the reconstructions. (3) For some concepts, lexical replacement by non-cognate words took place between Classical and Vulgar Latin, such as *via* → *strata* “path”,[15] *ignis* → *focus* “fire” (Meyer-Lübke 1935: 293), or *iecur* → *ficatum* “liver” (Herman 2000: 106). Again, this is reflected in the reconstruction.

On the negative side, the reconstructions occasionally reflect sound changes that only took place in the Western Romania, such as the voicing of plosives between vowels (Herman 2000: 46), as in `figat` “liver” (← *ficatum*).

[14] ASJP does not contain a word list for Vulgar Latin, so it is not possible to do a precise segment-by segment comparison.

[15] Latin makes a semantic distinction between *via* for unpaved and *strata* for paved roads; cf. (Meyer-Lübke 1935: 685)

Table 7: Reconstructions for Proto-Romance.

Concept	Latin	Reconstruction
blood	saNgw~is	saNg
bone	os	os
breast	pektus, mama	pet
come	wenire	venir
die	mori	murir
dog	kanis	kan
drink	bibere	beb3r
ear	auris	oreL3
eye	okulus	okyu
fire	iNnis	fok
fish	piskis	peS
full	plenus	plen
hand	manus	man
hear	audire	sentir
horn	kornu	korn3
I	ego	iy3
knee	genu	Z3nuL
leaf	foly~u*	foLa
liver	yekur	figat
louse	pedikulus	pidoko
mountain	mons	munta5a
name	nomen	nom
new	nowus	novo
night	noks	note
nose	nasus	nas
one	unus	unu
path	viya	strada
person	persona, homo	persona
see	widere	veder
skin	kutis	pel
star	stela	stela
stone	lapis	pEtra
sun	sol	sol
tongue	liNgw~E	liNga
tooth	dens	dEnt
tree	arbor	arbur
two	duo	dos
water	akw~a	akwa
we	nos	nos
you	tu	tu

Let me conclude this section with some reflections on how the reconstructions were obtained and how this relates to the comparative method.

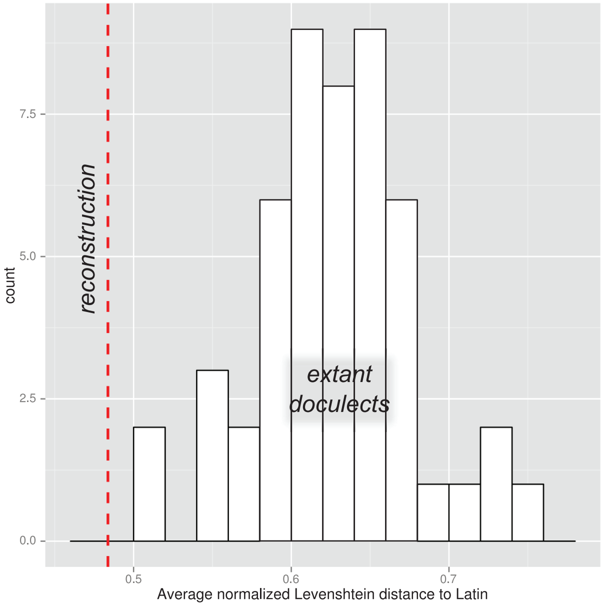

Figure 6: Average normalized Levenshtein distance to Latin words: reconstruction (dashed line) and extant Romance doculects (white bars).

A major difference to the traditional approach is the stochastic nature of the workflow sketched here. Both phylogenetic inference and ancestral state reconstruction are based on probabilities rather than categorical decisions. The results shown in Table 7 propose a unique reconstruction for each concept, but it would be a minor modification of the workflow only to derive a probability distribution over reconstructions instead. This probabilistic approach is arguably an advantage since it allows to utilize uncertain and inconclusive information while taking this uncertainty properly into account.

Another major difference concerns the multiple independence assumptions implicit in the probabilistic model sketched in Subsection 3.4. The likelihood of a phylogeny is the product of its likelihoods for the individual characters. This amounts to the assumptions that the characters are mutually stochastically independent.

The characters used here (and generally in computational phylogenetics as applied to historical linguistics) are mutually dependent in manifold ways though. For instance, the loss of a cognate class makes it more likely that the affected lineage will acquire another cognate class for the same semantic slot and vice versa.

This problem is even more severe for phonetic change. Since the work of the Neogrammarians in the nineteenth century, it is recognized that many sound changes are *regular*, i.e. they apply to all instances of a certain sound (given contextual conditions) throughout the lexicon. Furthermore, both regular and irregular sound changes are usually dependent on their syntagmatic phonetic context, and sometimes on the paradigmatic context within inflectional paradigms as well. Bouchard-Côté et al. (2013) and Hruschka et al. (2015) propose more sophisticated probabilistic models of language change than the one used here to take these dependencies into account.[16]

Last but not least, the treatment of borrowing (and language contact in general) is an unsolved problem for computational historical linguistics. Automatic cognate clustering does not distinguish between genuine cognates (related via unbroken chains of vertical descent) and (descendants of) loanwords. This introduces a potential bias for phylogenetic inference and ancestral state reconstruction, since borrowed items might be misconstrued as shared retentions.

4 Conclusion

This article gives a brief sketch of the state of the art in computational historical linguistics, a relatively young subfield at the interface between historical linguistics, computational linguistics and computational biology. The case study discussed in the previous section serves to illustrate some of the major research topics in this domain: identification of genetic relationships between languages, phylogenetic inference, automatic cognate detection and ancestral state recognition. These concern the core issues of the field; the information obtained by these methods is suitable to address questions of greater generality, pertaining to general patterns of language change as well as the relationship between the linguistic and non-linguistic history of specific human populations.

Acknowledgment: I thank Johann-Mattis List, Wiltrud Mihatsch, the co-authors of this issue Anne Breitbarth, Ailis Cournane and George Walkden, as well as the editor Hans-Martin Gärtner, for helpful feedback on draft versions of this article.

Funding: This research was supported by the DFG-Centre for Advanced Studies in the Humanities *Words, Bones, Genes, Tools* (DFG-KFG 2237), which is gratefully acknowledged.

[16] So far, these models have only been tested on one language family each (Austronesian and Turkic respectively), and the algorithmic tools have not been released.

References

Anthony, D. W. 2010. *The horse, the wheel, and language: How Bronze-Age riders from the Eurasian steppes shaped the modern world*. Princeton: PUB.

Atkinson, Q. D. & R. Gray. 2005. Curious parallels and curious connections — phylogenetic thinking in biology and historical linguistics. *Systematic Biology* 54(4). 513–526.

Atkinson, Q. D., A. Meade, C. Venditti, S. J. Greenhill & M. Pagel. 2008. Languages evolve in punctuational bursts. *Science* 319(5863). 588–588.

Baxter, W. H. & A. Manaster Ramer. 2000. Beyond lumping and splitting. Probabilistic issues in historical linguistics. In C. Renfrew et al. (eds.), *Time depth in historical linguistics*, vol. 1, 167–188. Cambridge: McDonald Institute for Archaeological Research.

Bergsma, S. & G. Kondrak. 2007. Multilingual cognate identification using integer linear programming. In *Proceedings of the RANLP Workshop*, 656–663.

Bouchard-Côté, A., D. Hall, T. L. Griffiths & D. Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences* 36(2). 141–150.

Bouckaert, R. et al. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097). 957–960.

Bouckaert, R. et al. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 10(4). e1003537.

Brown, C. H., E. Holman & S. Wichmann. 2013. Sound correspondences in the world's languages. *Language* 89(1). 4–29.

Campbell, L. 2013. *Historical linguistics. An introduction*. Edinburgh: EUB.

Chen, M.-H., L. Kuo & P. O. Lewis. 2014. *Bayesian phylogenetics. Methods, algorithms and applications*. Abingdon: CRC Press.

Covington, M. A. 1996. An algorithm to align words for historical comparison. *Computational Linguistics* 22(4). 481–496.

Csardi, G. & T. Nepusz. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* 1695(5). 1–9.

Desper, R. & O. Gascuel. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of computational biology* 9(5). 687–705.

Dolgopolsky, A. B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia. In V. V. Shevoroshkin (ed.), *Typology, relationship and time: A collection of papers on language change and relationship by Soviet linguists*, 27–50. Ann Arbor: Karoma Publisher.

Dunn, M., S. J. Greenhill, S. Levinson & R. D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473(7345). 79–82.

Durbin, R., S. R. Eddy, A. Krogh & G. Mitchison. 1989. *Biological Sequence Analysis*. Cambridge, UK: CUP.

Dyen, I., J. B. Kruskal & P. Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5). 1–132.

Edwards, A. W. F. & L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. In V. H. Heywood & J. R. McNeill (eds.), *Phenetic and phylogenetic classification*, 67–76. London: Systematics Association Publisher.

Embleton, S. M. 1986. *Statistics in historical linguistics*. Bochum: Brockmeyer.

Ewens, W. & G. Grant. 2005. *Statistical methods in bioinformatics: An introduction*. New York: Springer.

Fitch, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology* 20(4). 406–416.

Fitch, W. M. & E. Margoliash. 1967. Construction of phylogenetic trees. *Science* 155(3760). 279–284.

François, A. 2015. Trees, waves and linkages: Models of language diversification. In C. Bowern & B. Evans (eds.), *The Routledge handbook of historical linguistics*, 179–207. Abingdon: Routledge.

Gascuel, O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* 14(7). 685–695.

Gray, R. D. & Q. D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(27). 435–439.

Gray, R. D. & F. M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405(6790). 1052–1055.

Gray, R. D., A. J. Drummond & S. J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913). 479–483.

Greenhill, S. J., R. Blust & R. D. Gray. 2008. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4. 271–283.

Haak, W. et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522(7555). 207–211.

Hall, D. & D. Klein. 2010. Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1030–1039. ACL.

Hammarström, H., R. Forkel, M. Haspelmath & S. Bank. 2016. *Glottolog 2.7*. Max Planck Institute for the Science of Human History, Jena. Available online at http://glottolog.org (accessed 29 January 2017).

Haspelmath, M., M. S. Dryer, D. Gil & B. Comrie. 2008. The World Atlas of Language Structures online. Munich: Max Planck Digital Library. http://wals.info/.

Hauer, B. & G. Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of the 5th International Joint NLP conference*, 865–873.

Heggarty, P., W. Maguire & A. McMahon. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1559). 3829–3843.

Herman, J. 2000. *Vulgar Latin*. University Park, PA: The Pennsylvania State University Press.

Hogeweg, P. & B. Hesper. 1984. The alignment of sets of sequences and the construction of phyletic trees: An integrated method. *Journal of molecular evolution* 20(2). 175–186.

Hruschka, D. J., S. Branford, E. D. Smitch, J. Wilkins, A. Meade, M. Pagel & T. Bhattachary. 2015 Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology* 25(1). 1–9.

Jäger, G. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* 3(2). 245–291.

Jäger, G. & J.-M. List. 2017. Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists. *Language Dynamics and Change* 8(1). 22–54.

Jäger, G. & P. Sofroniev. 2016. Automatic cognate classification with a Support Vector Machine. In S. Dipper et al. (eds.), *Proceedings of the 13th Conference on Natural Language Processing*, 128–134. Bochum: RUB.

Jäger, G., J.-M. List & P. 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. ACL.

Kay, M. 1964. *The logic of cognate recognition in historical linguistics*. Santa Monica, CA: Rand Corporation.

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

Kondrak, G. 2002. *Algorithms for language reconstruction*. University of Toronto PhD thesis.

Kooperberg, C. 2016. Package 'logspline'. https://cran.r-project.org/web/packages/logspline/index.html. version 2.1.9.

Kroonen, G. 2013. *Etymological dictionary of Proto-Germanic*. Leiden, Boston: Brill.

Lewis, M. P., G. F. Simons & C. D. Fennig (eds.). 2016. *Ethnologue: Languages of the world*, 9th edn. Dallas, Texas: SIL International.

List, J.-M. 2012. Lexstat: Automatic detection of cognates in multilingual wordlists. In M. Butt & J. Prokić (eds.), *Proceedings of LINGVIS & UNCLH, Workshop at EACL 2012*, 117–125, Avignon.

List, J.-M. 2014. *Sequence comparison in historical linguistics*. Düsseldorf: DUP.

Lowe, J. B. & M. Mazaudon. 1994. The reconstruction engine: A computer implementation of the comparative method. *Computational Linguistics* 20(3). 381–417.

McMahon, A. & R. McMahon. 2005. *Language classification by numbers*. Oxford: OUP.

McMahon, A. & R. McMahon. 2006 Why linguists don't do dates: Evidence from Indo-European and Australian languages. In P. Forster & C. Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 153–160. Cambridge, UK: McDonald Institute for Archaeological Research.

Meillet, A. 1954. *La méthode comparative en linguistique historique*. Paris: Honoré Champion.

Meyer-Lübke, W. 1935. *Romanisches etymologisches Wörterbuch*. Heidelberg: Carl Winters Universitätsbuchhandlung. 3. Auflage.

Needleman, S. B. & C. 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. 443–453.

Nguyen, L.-T., H. A. Schmidt, A. von Haeseler & B. Q. Minh. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32(1). 268–274.

Notredame, C., D. G. Higgins & J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* 302(1). 205–217.

Oakes, M. P. 2000. Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics* 7(3). 233–243.

Pagel, M., Q. D. Atkinson & A. Meade. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163). 717–720.

Pagel, M., Q. D. Atkinson, A. S. Calude & A. Meade. 2013. Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences* 110(21). 8471–8476.

Pietrusewsky, M. 2008. Craniometric variation in Southeast Asia and neighboring regions: a multivariate analysis of cranial measurements. *Human Evolution* 23(1–2). 49–86.

Raghavan, U. N., R. Albert & S. Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* 76(3). 036106.

Rama, T. 2013. Phonotactic diversity predicts the time depth of the world's language families. *PLoS ONE* 8(5). e63238.

Rama, T. 2015. Automatic cognate identification with gap-weighted string subsequences. In *Proceedings of the North American Association for Computational Linguistics*, 1227–1231. ACL.

Renfrew, C. 1987. *Archaeology and language: The puzzle of Indo-European origins*. Cambridge, UK: CUP.

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

Ringe, D. A., T. Warnow & A. Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1). 59–129.

Ronquist, F. & J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12). 1572–1574.

Ross, M. & M. Durie. 1996. Introduction. In Mark Durie & Malcolm Ross (eds.), *The comparative method reviewed. Regularity and irregularity in language change*, 3–38. Oxford: OUP.

Saitou, N. & M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4). 406–425.

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9). 1312–1313.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96(4). 452–463.

Swadesh, M. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21. 121–137.

Turchin, P., I. Peiros & M. Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship* 3. 117–126.

Weiss, M. 2015. The comparative method. In C. Bowern & B. Evans (eds.), *The Routledge handbook of historical linguistics*, 119–121. London: Routledge.

Wichmann, S., E. W. Holman & C. H. Brown. 2016. The ASJP database (version 17). http://asjp.clld.org/.

Article note: All code and data used and produced when conducting the case study in Section 3 are available for download and inspection from https://github.com/gerhardJaeger/protoRomance.