3 Computational Approaches to Linguistic Chronology and Subgrouping

Dariusz Piwowarczyk

3.1 Computational Approaches to Historical Linguistics

Computational approaches to historical linguistics can be roughly divided into those pertaining to language classification (e.g. Ringe et al. 2002), chronology (e.g. Gray & Atkinson 2003), cognate detection (e.g. Kondrak 2002), comparative reconstruction (e.g. Hewson 1974) and the simulation of phonological (Baker 2008) and analogical change (Skousen 1989).[1] They usually involve quantitative methods to calculate the relationship between the languages or to date the chronology of their split from the proto-language, as well as algorithms to align cognates, reconstruct proto-forms or simulate sound changes (including artificial neural networks for analogical changes).

Although the use of quantitative methods is not new in contemporary linguistics – they were already being used in the 1930s – the application of computational methods in historical linguistics, like those used in evolutionary biology, represents a novel approach which is gaining adherents but is mostly regarded as problematic by traditional historical linguists. This is partly because of the history of quantitative approaches, which included methods such as glottochronology and lexicostatistics that are now found to be largely unreliable (cf. Bergsland & Vogt 1962) or controversial at the very

This work has been written under the research project financed by the National Science Centre (Poland) decision number: 2018/02/X/HS2/03669. I am grateful to Thomas Olander, Don Ringe and Michael Weiss for comments on the earlier version and to Pete Westbrook for correcting my English. Needless to add, I am solely responsible for errors and mistakes.

[1] Cf. Sims-Williams 2018 for a recent short overview of past approaches. Computational approaches to historical linguistics have been expanding ever since the early 2000s. Nowadays, there is an enormous outgrowth in works dealing with computational historical linguistics in all its aspects (language classification, cognate alignment, sound change simulation, analogical change simulation). For an overview of the most recent approaches, see Dunn 2015 and Jäger 2018. For a short overview and assessment of quantitative approaches to historical linguistics, see the discussion in the third edition of Lyle Campbell's handbook (Campbell 2013: 447–92).

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

least (cf. Hoijer 1954). Additionally, many linguists approached such methods with extreme caution because they involved handling and converting the data to a machine-readable format, using statistical algorithms which seemed to generate a “black-box” effect rather than explain the results, and comparing language development to the replacement of genetic material in evolutionary biology. Compounding the problem was the fact that many of these early computational approaches were actually implemented by computational biologists rather than linguists.

Even though the methods and the data used have often not been applied very methodically or carefully (especially in the early works dealing with linguistic classification and chronology), there is no doubt that computational methods can be useful in linguistics, simply because computers can analyse masses of data in a short space of time and without errors. This might not be helpful in all aspects of historical linguistics, but it will certainly make it easier to check, for example, the coherence of a hypothesis put forward, to test competing hypotheses or the results of the application of a sound change etc. However, it has to be borne in mind that:

1. The computer is only a tool and the results yielded will always ultimately depend on the quality of the algorithm, the input data and how they are converted into a machine-readable format (together with all the judgements made by the researchers at this point).
2. Research results from using computational methods have to be interpreted, and the method itself is usually meant as a supplement to traditional methods, not a replacement.

In this chapter, I will outline the history of computational approaches to historical linguistics, concentrating on those concerning language classification and chronology, and then describe the three major contrasting approaches to the classification and chronology of the Indo-European languages. Following this, I will present the method used for the computational replication of sound change, which is a useful tool for testing the relative chronology of sound changes and may have some bearing on the grouping of the Indo-European languages. I conclude with some perspectives for future work.

3.2 Sources and Origins

Although most of the contemporary computational approaches to language classification and chronology stem directly from the methods used in evolutionary biology (cf. Dunn 2015), they seem to have their indirect roots in the earlier quantitative approaches which used statistical and mathematical methods to calculate different aspects of language change and comparison.[2]

[2] Only the main approaches are outlined here. For details on the different statistical approaches and their history, I refer the reader to the overview of the use of statistics in historical linguistics by

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

This clearly stems from the ongoing search for more objective bases to support the traditional arguments from historical linguistics, which were often seen as subjective as they depended largely on the assumptions of the scholars who performed the research.

One of the earliest quantitative approaches was a method devised by the Polish anthropologist Jan Czekanowski (1927), who tried to calculate the similarities between the Indo-European languages and present the results in a numerical way. Czekanowski approached the Polish linguist Jerzy Kuryłowicz for a list of characteristic binary features which could be used in distinguishing the different subgroups of Indo-European (at first twenty, then twenty-two) and proceeded to create the two-feature contingency tables in which those features were counted for every language that was being compared. Then he used Pearson's tetrachoric correlation formula known from statistics. This allowed him to present a distance matrix of the Indo-European languages. His approach made its way to the United States through his student in anthropology, Stanisław Klimek, who came to study with Alfred Kroeber in the 1930s. Czekanowski's method was adopted by both Alfred Kroeber and Charles Douglas Chrétien (1937), who tried to count the similarities between the Indo-European languages using a broader range of features (seventy-four in total) taken from Meillet's monograph on the Indo-European dialects (1922). However, their results were criticised (cf. Safarewicz 1948) for being biased from the very start because of the use of data based on a work which itself intended to prove the groupings of, for example, Italo-Celtic (Meillet 1922). Probably the most famous approach was championed in the 1950s by Morris Swadesh (1952) in the form of glottochronology. Although promising at first, it was harshly criticised for working on the initial assumption that there is a constant rate of change in languages. Embleton (1986) tried to further enhance the computerised methods of lexicostatistics and glottochronology and concluded that

> for the traditional methods as well as the statistical methods the reconstruction of the topology of the tree is more accurate than the assignment of dates. Reliable dating information is more likely to come from historical or archeological sources, although the statistical methods can provide some provisional estimates. (Embleton 1986: 169–70)

3.3 Computational Approaches to Language Classification and Chronology

As mentioned above, most of the contemporary computational approaches to language classification and chronology stem from the methods used in

Sheila Embleton (1986) and the monograph on word lists and lexicostatistical approaches to language comparison by Brett Kessler (2001). For an overview and a summary assessment of both glottochronology and early lexicostatistics, see Tischler (1973).

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

biological sciences.[3] The computational approaches used in evolutionary biology were applied to linguistics in the late 1990s and began to be used in the early 2000s. They usually use statistical Bayesian inference to infer phylogenies. This kind of work has become very popular and has already been applied to different families of languages.

An interesting distance-based approach was pioneered by Søren Wichmann and his team (2018) in the Automated Similarity Judgement Program and the corresponding ASJP database (Wichmann et al. 2018). The database includes a list of forty basic words for more than 5,000 languages and can be used, for example, to date when the languages in one family split away from each other. Because it uses the Levenshtein distance and lexical data, it is often regarded sceptically by linguists (cf. Greenhill 2011).[4] Additionally, there are some errors in the database itself. In the word list for Latin, for example, there is no vowel length present, and the words are transcribed inconsistently: *wenire* 'to go' is transcribed with /w/ whereas *viya* 'road' is transcribed with /v/.

One of the most controversial aspects of computational (or more accurately statistical) approaches to language classification and chronology is the fact that they are heavily based (often even exclusively based) on lexical data. In contrast, the standard procedure in traditional historical linguistics is the analysis of phonological and morphological features, and this is probably the main reason that many traditional historical linguists are generally very sceptical about using computational approaches for language classification and chronology. As pointed out by Gerhard Jäger and Johann-Mattis List in their recent comparison of traditional and computational methods, the crucial difference between the classical comparative method and the approaches adopted by computational historical linguistics is that "the comparative method strives to reconstruct the *true* history of languages in their entirety while statistical approaches search for *probable* or at least *useful* models of the observed patterns in some well-defined partial range of data".[5]

3.4 Application to Indo-European Studies

Apart from the lexicostatistical approach employed by Dyen et al. (1992) using the 200-word Swadesh lists for ninety-five languages and assuming a similar rate of change in all languages, probably the most famous computational

[3] For an in-depth overview of the methods, I refer the reader to Nichols & Warnow (2008) and, for a discussion of more recent studies in the area, to Dunn 2015.

[4] "The Levenshtein distance is a simple distance metric derived from the number of edit operations needed to transform one string into another" (Greenhill 2011: 689).

[5] Gerhard Jäger & Johann-Mattis List, 2019, Statistical and computational elaborations of the classical comparative method (unpublished manuscript), https://bit.ly/3yVktOs (accessed 20 February 2020): 30.

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

classification of Indo-European languages was developed in 2002 by a team of experts combining linguistics (Don Ringe, Ann Taylor) and computer science (Tandy Warnow) in the project on *Computational phylogenetics in historical linguistics* (with contributions from statistician Steven Evans).[6] Using 22 phonological, 13 morphological and 259 lexical features as coded characters, they were able to produce a tree with a "perfect phylogeny" algorithm that tracked the branching of twenty-four ancient and medieval Indo-European languages. However, the phylogeny was not quite perfect since the position of Germanic could not be determined. As it turned out from subsequent work, which included language contact (Nakleh, Ringe & Warnow 2005), this was due to the fact that Germanic was apparently in contact with the other branches and therefore did not fit the "perfect phylogeny".

Probably the most controversial computational approach to the subgrouping and chronology of the Indo-European languages was that adopted by Russell Gray and Quentin Atkinson (2003). In their research, Gray and Atkinson used the word lists of basic vocabulary for eighty-seven Indo-European languages compiled by Dyen et al. (1992) along with Dyen et al.'s cognancy judgements and applied Bayesian inference to establish the dates for linguistic divergence of the languages analysed. They employed the algorithms for estimating the divergence time of DNA from evolutionary biology calibrated to the dates of the languages' known split times. Using this technique, they were able to generate a tree in which the estimated dates of divergence of the particular groups of Indo-European languages were essentially in line with Colin Renfrew's theory on the Anatolian origin of Indo-European languages (Renfrew 1987). The method was further expanded using phylogeographic approaches by Bouckaert et al. (2012), with the results also pointing to an Anatolian origin.

The work of Gray & Atkinson (2003) and Bouckaert et al. (2012) was challenged by a team from the University of California, Berkeley (Chang et al. 2015). They tried to use the same method but with the addition of ancestry constraints, i.e. information relating to the fact that Latin is the parent language of Romance etc. Their research indicated that the chronology of the Indo-European splits was significantly shorter than previously thought and roughly in accordance with the dating of the so-called steppe hypothesis in archaeology as regards the homeland of the speakers of Proto-Indo-European. Although the authors claim that "the agreement between our findings and the independent results of other lines of research confirms the reliability of statistical inference

[6] I will outline here only the three main applications used for the classification, subgrouping and chronology of Indo-European languages as represented in Ringe et al. 2002, Gray & Atkinson 2003 and Chang et al. 2015. I refer the reader to Pereltsvaig & Lewis 2017 and McMahon & McMahon 2006 for a more in-depth analysis. A summary and assessment of most of these approaches was also given by Ringe (2017b: 67–71).

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

of reconstructed chronologies" (Chang et al. 2015: 194), there are problems inherent in the Bayesian inference itself (cf. Ringe in Chapter 4 and the comments by Nichols & Warnow 2008: 785 on Bayesian methods).

At present, it is difficult to say whether computational methods can give us a reliable chronology of the splits of the individual branches.[7] It is clear, though, that they can provide us with a reliable indication of how the trees (topology) branched, as the work by Ringe et al. (2002) has shown. It is also striking that most of the approaches identify the traditionally assumed subgroups (Indo-Iranian, Balto-Slavic, Italo-Celtic, Graeco-Armenian). Even with careful calibration, perhaps the best we can hope for is a very rough estimate. However, even this should be corroborated by linguistic, archaeological, genetic and historical data.

3.5 Computational Replication of Sound Change (Computerised Forwards Reconstruction)

Another approach which could yield promising results in the reconstruction of the Indo-European family tree, especially with regard to the relative chronology of sound changes, is the replication of sound change or computerised forwards reconstruction (cf. Sims-Williams 2018).[8] The procedure of reconstructing forwards is not unknown in traditional historical linguistics. For example, Calvert Watkins (1962: 5) mentions it in his monograph on the Indo-European origins of the Celtic verb, where he employs the method to see how a Proto-Indo-European form would be regularly continued in Celtic. The method was also adopted by Ives Goddard (1998: 183) in his analysis of the Arapaho historical phonology, which, like the phonology of the insular Celtic languages, underwent some radical changes.

The aim of such an approach, enhanced by using a computer, is to employ an algorithm that reads the given input data (proto-forms), makes the appropriate changes based on the changes that are usually assumed to have taken place in the development of the particular languages (regular sound changes) and generates the output which can be either manually or automatically compared with the actually attested forms. The purpose of this approach is to test the regularity of sound changes and their relative chronology. If the generated form is exactly the same as the attested one, then the relative chronology is assumed to be correct.

However, only two programs known to the author have tried to replicate sound changes comprehensively by taking into account most of the changes (Maniet 1985; Hartman 2003). Most of the other programs have only used

[7] See also Chapter 4 by Ringe on the limits of computational cladistics.
[8] This is also sometimes called "historical derivation"; cf. Kondrak 2002: 12–15.

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

fragmentary data and applied only the main sound changes. I will list and discuss the most important ones below.[9]

The first application of historical derivation was a program by Raoul Smith (1969) which applied twenty-one regular rules to 650 Proto-Indo-European reconstructed lexemes as taken from an etymological dictionary and, through the application of those rules, derived the modern Russian forms from the Proto-Indo-European ones. However, only nine lexemes were derived regularly because too few sound-change rules were applied.

Further attempts were made by Burton-Hunter (1976) to generate Old French forms from their Latin sources and by Eastlack (1977), to produce Ibero-Romance from its Latin source. Both programs yielded a high percentage of expected forms since they took into account a larger number of rules within a shorter period of time (Latin to Old French or Old Spanish).

Probably one of the most comprehensive computational sound-change replication programs was the one devised by Albert Maniet (1985). He simulated the changes from Proto-Indo-European to Latin (252 rules) on a corpus of approximately 15,000 words from Plautus. However, his research went largely unnoticed by both linguists and computer scientists, and the program is, unfortunately, no longer accessible today.

In more recent times, in his thesis on cognate alignment and reconstruction, Kondrak (2002: 141–3) included a short appendix on a Perl program which generated Polish forms from their Proto-Slavic sources. From the 626 lexemes taken into account, 72.5 per cent were regular.

Hartman (2003) has been developing a program since the 1980s which simulates the sound changes from Latin to Spanish (with approximately 122 rules and 1,806 coded vocabulary items) using sets of distinctive features (from Chomsky & Halle 1968) coded as binary strings rather than the usual string substitution as in most programs developed so far. As input, the etymon is "fed" into the computer via the keyboard or from a file. Then the individual letters are translated into sets of features, and changes are applied to the features in accordance with the programmed rules and their relative chronology. Finally, the features are translated back into characters and displayed using the International Phonetic Alphabet.

Hartman also enabled the program itself to be used to model sound changes in other languages. Recently, the hard-coded version with the Spanish model was made available on-line, and a new working version of the program was presented (Hartman 2018) but with no significant modifications. The earlier version of Hartman's program was used by Towhid bin Muzzafar to simulate the changes from Proto-Algonquian to Shawnee (bin Muzzafar 1997). Apart

[9] I also refer the reader to the enumeration and short descriptions of the programs by Sims-Williams in his overview (2018).

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

from some additional examples that he was able to find for Shawnee and minor improvements in the relative chronology of changes, he pointed out an important aspect of the computational replication of sound change which probably constitutes the main obstacle to this kind of work:

> It has been known that if diachronic sound change is regular, then it must be possible to demonstrate the regularity of sound change in computer models. But very few have actually ventured to take historical sound change rules from textbooks of well studied languages and develop a working computer model. And anyone who HAS ventured into this territory has quickly realized that there is a world of difference between the rules as they are written in standard linguistic notation and as they need to be written in computer models. (bin Muzzafar 1997: 73–4)

Some more recent programs, apart from the ones developed by people to create fictitious languages (so-called "Conlangers"), include the simulation of Spanish from Latin by Marcel Schmuki ("ETYMO", 2001), Proto-Germanic from Proto-Indo-European by Brett Kessler ("Derive", 2004), Gothic from Proto-Indo-European by Roland Mittmann (2009) and the "Sound change transducers" by Amir Zeldes (2008).

Sims-Williams tried some computational replication of sound change in Celtic historical phonology just by using the "Find & Replace" function in the word processor (Sims-Williams 2018: 562). By applying forty-three sound changes on the material of 159 selected Common Celtic forms, he was able to find amendments in the relative chronology of changes and identify usually overlooked Celtic cognates from a Proto-Indo-European root. He further argued that with modern programs, this kind of research could be pursued and expanded significantly so that it might include complete Proto-Indo-European reconstructions as input to find new etymologies, correct existing ones and check the relative chronology of sound changes (Sims-Williams 2018: 564). I agree with the author in principle, although I think implementing such an expanded approach is quite complicated, because of several key factors:

1. circularity of the method – there are competing hypotheses on what the Proto-Indo-European reconstructed forms should look like, and the applied sound changes can be used to fit the proto-form
2. problems in "translating" the sound changes into computational notation
3. problems in coding the forms and characters as the input for the program
4. lack of high-quality computational databases for Indo-European languages and Proto-Indo-European which would also include verbal and nominal paradigms
5. lack of monographs and works that comprehensively present the complete relative chronology of changes from Proto-Indo-European to the attested languages

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

6. not applying morphological changes which are usually assumed to have been very numerous in paradigms and which will inevitably blur the regular outcomes from simulations, especially if the simulation has a very large time span (e.g. from Proto-Indo-European to Latin etc.).

To give just one example, if we take the Proto-Indo-European reconstructed form for ‘foot’ (cf. (Wodtko et al. 2008: 530 n. 2) and apply the sound changes that occurred in Attic Greek (cf. Liesner 2015: 110–15), we will get the following result:

nom.sg.	**pód-s*	>	**pots*	>	**pos* (preserved in cpd. *τρίπος* ‘tripod’)
gen.sg.	**péd-s*	>	**pets*	>	**pes*

Apart from the fact that some forms do not match the attested ones because of not being included in the morphological changes, we will encounter problems with the reconstruction of the proto-form, since different scholars propose competing forms: nom.sg. **pṓds*, gen.sg. **pedés* (Ringe 2017a: 59) or nom.sg. **pṓds*, gen.sg. **pedós* (Clackson 2007: 72). There are also problems with the assumed sound changes (cf. Szemerényi 1996: 116 for the view that **póds* developed into **póss* and Greek **pṓs*).

3.6 Potential Ways to Enhance the Computational Replication of Sound Change

In this section, I will try to address some of the problems involved in computational replication of sound change.

3.6.1 Circularity and the Assumption of Different Reconstructed Input Forms

Probably the easiest way to avoid circularity would be to test the various competing hypotheses, which would be fairly easy if the program were interactive and allowed changes to the rules and the input forms. However, the correct form might still be different from any of those proposed so far. Therefore, what remains is to try to use an algorithm that reconstructs the proto-form or the one that would infer the changes (based on the ones that are known e.g. from Latin to Romance) rather than project them mechanically in a replicatory manner (cf. Anderson, List & Tresoldi 2018). The exact mechanism of this approach has not been fully presented yet, but if it turns out to be successful, it could prove an additional help to our understanding of linguistic changes.

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

3.6.2 Problems in "Translating" the Sound Changes into Computational Notation

This is one of the most important obstacles, since the linguistic changes cannot simply be used in computational algorithms but have to be "translated" into computational terms. That is, either as simple string substitutions as in most programs or with the use of distinctive features through a binary matrix or parsing.[10] Either way, only simple sound changes such as Latin rhotacism can be easily coded in terms of string substitutions. Problems occur with changes that operate only in certain positions in the word (since the computer does not know what a syllable is, so this requires additional coding) or, even worse, changes that depend on factors outside of phonetics and phonology (e.g. the apocope of final *-*i* in Latin, which occurs in verbs and adverbs but not in nouns and could be triggered by the final position of the verb in the sentence, cf. Hock 2012, if correct).

3.6.3 Problems in Coding the Forms and Characters as the Input for the Program

Apart from translating sound changes into computational terms, the input also has to be coded in such a way as to encompass all the necessary features depending on the required changes (accent, prosody, co-articulation etc.). This, along with the mutual compatibility of the data, was recently addressed by List (2017), who proposed a universal format for coding etymological data.

3.6.4 Lack of High-Quality Databases for Indo-European Languages and Proto-Indo-European

This is also an important problem since the outcome of a computer simulation will essentially depend on the quality of the input and the programmed rules. There are hardly any reliable databases for Indo-European studies, but there are some recent projects (see Noyer 2016; Barnett & Macdonald 2018), which, once completed and made available, should remedy the situation to some extent.

[10] A string is understood as either a letter or a number, in this case only a letter. An example of a programmed sound change would be Latin rhotacism, where the string substitution would be programmed as follows: change every /VsV/ sequence into /VrV/ in all forms in the database. The vowel (V) would be defined as either /a(:)/, long or short respectively, /e(:)/, /i(:)/, /u(:)/, /o(:)/. The algorithm would then proceed to change every VsV sequence into VrV automatically and without exceptions, thus replicating what is usually thought of as an example of regular sound change.

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

3.6.5 Lack of Monographs with the Complete Relative Chronology of Changes

This is more of a problem for Indo-European studies in general than it is for the computational replication of sound change. There seems to be a lack of comprehensive works which present the complete (even hypothesised) relative chronology of changes. In most publications, only the chronology of main changes is given (even in e.g. McCone 1996) or the ones that are relevant to the discussion. This has been improving in recent years (cf. Matasović 2005; Olander 2015: 46–67), but much work in this area remains to be done.

3.6.6 Not Applying Analogical Changes

Whereas it is clear how to simulate regular sound changes, analogical changes are inherently irregular and occur quite unpredictably (see Olander 2015: 46). If that is so, it will be problematic to code them appropriately, other than as a substitution of the whole word within the chronology of changes: e.g. gen.sg. **pód-s* → **pód-os* with this change occurring only in this specific form. This can create problems if there are two similar forms in the paradigms and analogical change occurs in only one of them. In that event, morphological annotation of the forms would be necessary to avoid such situations.

3.7 The Potential Use of Relative Chronology of Sound Changes in Subgrouping

In an article devoted to the position of West Germanic, Don Ringe observed that

> the chronology of changes serves two purposes. On the one hand, languages are much less likely to have undergone innovations in the same order independently by chance. On the other hand, a sequence of changes should require more time to go to completion than a similar set of unrelated changes, thus ensuring that the period of linguistic unity demonstrated by the shared changes continued for a significant period of time. (Ringe 2012: 33)[11]

If that is so, then it would be possible to use the computational replication of sound change in this area as well, depending on the quality and the availability of the data as discussed below.[12]

[11] Cf. the similar arguments made by Matasović 2005 for Balto-Slavic.

[12] For an in-depth discussion of the particular subgroups, I refer the reader to the respective chapters in this volume.

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

3.7.1 Indo-Iranian

The relative chronology of the main Indo-Iranian sound changes along with the approximate reconstruction of Proto-Indo-Iranian seem more or less established: Lubotsky (2018: 1885–6) gives a list of ten consecutive sound changes common to Indic and Iranian. Difficulties in the computational simulation include differing opinions between scholars on the place of Bartholomae’s Law in the relative chronology of changes (was it also a Proto-Indo-European process or solely Indo-Iranian or perhaps two independent changes?) or the exact conditioning of Brugmann’s Law and reconstruction of the proto-forms accordingly (e.g. the Proto-Indo-European reconstruction *_h_$_3$_éu̯is_ or *_h_$_2$_óu̯is_ ‘sheep’ depending on the assumptions made about the conditioning of Brugmann’s Law and the absence of its operation in this word either due to the full grade of the ablaut or analogical change).

3.7.2 Balto-Slavic

There is considerable discussion about the relative chronology of Slavic and Baltic sound changes, but, although there are differences between scholars on the details and the exact relative chronology (cf. Matasović 2005; Kortlandt 2008; Olander 2015), the main sound changes seem to be established, indicating the existence of a subgroup: Olander (2015: 47–53) lists eleven consecutive sound changes common to Balto-Slavic. Technical problems arise in the computational simulation with the “translation” of the changes in computerised terms concerning the Balto-Slavic accentuation and coding of the accent. Furthermore, there is a problem with the double reflex of Proto-Indo-European syllabic resonants in Balto-Slavic as either *_iR_ or *_uR_, since the exact conditioning has to be stated for the computational simulation.

3.7.3 Italo-Celtic

For Italo-Celtic, sound changes seem relatively less important than morphological changes (cf. de Vaan 2008: 7; Weiss 2020: 493–5). Weiss (2020: 207) gives a list of four consecutive sound changes that could be common to Italic and Celtic. There are also controversies concerning whether this stage existed at all – compare the arguments of Meiser (2003: 30–1) and the discussions by Schrijver (2006: 48–53) and also recently by Zair (2018). There is hardly any complete hypothesis on the reconstruction of Proto-Italo-Celtic (cf. Kortlandt 2007: 149–78) or even a balanced account of Proto-Italic (cf. van der Staaij 1995).

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

3.7.4 Graeco-Armenian

The relative chronology of changes usually postulated for Ancient Greek and Armenian does not seem to support a Graeco-Armenian subgroup (cf. Kim 2018). Mostly lexical items favour this grouping. It is also the weakest in the computational cladistic analysis by Ringe et al. 2002.

From the point of view of the chronology of sound changes, only the Indo-Iranian and Balto-Slavic subgroups appear to be real entities. Proto-Indo-Iranian and Proto-Balto-Slavic also have more or less established reconstructions. However, in order to be able to fully investigate the relative chronology of sound changes, it would be necessary to compile a comprehensive list, have the changes translated as closely as possible from linguistic to computational notation and use a high-quality database with more or less complete data.

Since it has long been a gold standard in historical linguistics that morphological innovations should be taken more seriously into account than phonological or lexical ones, in the next section, I will discuss the possibilities and perspectives of including morphological changes in the computational replication of sound change.

3.8 Perspectives on the Inclusion of Morphological Changes

It may be possible to expand the scope of the computational replication of sound change in such a way as to apply computationally generated sound changes along with morphological changes to the complete lexicon of the Proto-Indo-European language as it is reconstructed today in order to generate the main Indo-European languages. The program would apply the sound changes and the analogical changes in their relative chronology to the lexicon of Proto-Indo-European and generate output which in turn would be compared with the actual data relating to those languages. With this approach, the amount of regular sound change from a more or less complete lexicon would be uncovered along with the exact interferences causing irregularities – errors in the formulation, chronology or translation into computational terms of the programmed sound changes, borrowing and, especially, analogy.

This approach could potentially address a very direct and practical question of interest to every practising historical linguist: whether one analogical solution is more probable than another. The usual answer to this question depends on one’s view of the system of the language in which the change occurred. However, different scholars might view a certain language system (in its earlier or even reconstructed phases) differently and so pose different analogical explanations with their own models and motives. They will look for parallel developments and typologically similar changes in the material they are working on and, most importantly, in their previous experiences. We can deem one

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

solution of analogical remodelling as more plausible than another by providing other analogical solutions of exactly the same type, with similar models and motives along with an in-depth analysis of the synchronic situation. Warren Cowgill, in his work on universals in Indo-European diachronic morphology, noted that with regard to small-scale innovations "[a] sufficiently large collection of such individual changes, appropriately classified, should give linguists measure of the relative plausibility of different solutions for problems in historical grammar" (1966: 115). He continues:

> At present each linguist judges the plausibility of a newly proposed solution pretty much by what he happens to remember of the morphologic innovations which during his career he has been led, for one reason or another, to accept as plausible. A reasonably objective standard of plausibility should make it easier for historical linguists to agree on solutions for problems of historical morphology that at present are still disputed. (Cowgill 1966: 115–116)

Using a computational algorithm and an electronic database with word forms and the phonological and morphological changes which occurred in their development, it would be possible to create a virtually complete picture of the phonological and morphological system of the language at every stage of its development and to investigate any possible phonetic and analogical changes. Because the reconstruction of the proto-language and each of its stages of development remains hypothetical, its validity and accuracy can only be checked against the general typology of both synchronic language systems and types of diachronic changes along with the internal coherence of the system. Most notably, the compatibility of every single sound change can be checked against the hypothesis by applying it to the lexicon of the language.

Such an approach would allow us to formulate hypotheses concerning the relative chronology and tendencies of sound change and analogical levelling based on fairly complete empirical data. The results would confirm or challenge the existing theories on sound change and analogical remodelling and could form the basis for comprehensive historical grammars in the future which, with the expansion of integrated corpus linguistics, could encompass all corpora of texts from all periods of the documented language development. Such a large database would enable scholars to pursue further research in the area, allow the explicit discussion of competing hypotheses and serve as an educational tool. Additionally, the method itself could be applied to other language families, thus forming the basis for research on universal tendencies in language change. Moreover, it would break the so-called "handbook tradition" mentioned by Eichner (1992: 61), whereby a sound change is illustrated only by a handful of examples (usually the same in various historical grammars) and in order to find more of these, one has to consult an etymological dictionary.

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

In fact, a recent project carried out by Jouna Pyysalo (2017) has managed to achieve most of what is described above. With the use of computational simulation, the project aims to generate all the forms of the Indo-European languages from their reconstructed proto-forms. However, the author uses an idiosyncratic reconstruction of Proto-Indo-European which deviates significantly from the current *communis opinio*, i.e. with only one laryngeal, at the same time basing his argument against the classical laryngeal theory only on Anatolian data and completely ignoring the data relating to Ancient Greek (cf. Janhunen & Pyysalo 2018). If the input proto-forms are hard-coded so that they cannot be modified, this project will only serve to present the author's own views on the subject.

3.9 Perspectives on the Computational Methods

Czekanowski, Kroeber and Chrétien pursued an interesting way of handling language classification back in the days when such methods were far from popular or even acceptable. They were often very harshly criticised by linguists, to the extent that virtually nobody followed their lead to improve the method. Just as Czekanowski, Kroeber and Chrétien were pioneers in the use of statistics for language classification applied to Indo-European, so were Ringe et al. and Gray and Atkinson pioneers in the use of computational methods in the same area. Even though their work is relatively recent, a large amount of new research has been done in the field, which has become so popular that some scholars argue that historical linguistics appears to have taken something of a quantitative turn. Indeed, new methods are being implemented in an attempt to meet the standards of traditional historical linguistics in paying careful attention to the annotation of the data and to detail in general, while at the same time taking advantage of the replicability, robustness and formality of the computational approach. Progress is being made in modelling language characteristics and change using algorithms, and more attention is being given to making the programs and the data openly accessible, in a more or less standard format, easy to use by non-computational scientists and, importantly, annotated jointly by experts in the respective fields. It seems that only through combining qualitative and quantitative methods can further progress be made in the field of Indo-European linguistics, and the current thinking is that such progress is only possible if scholars from different disciplines contribute collectively.

The advantages of the computational approach are its speed, error-free processing of data and ability to handle large amounts of data. I would further argue that, thanks to the computational approach and the explicit presentation of the material, it would be much easier to compare different linguists' competing hypotheses, even for people from outside the exact field of

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

specialisation, thus making the whole enterprise of Indo-European linguistics easily accessible for interdisciplinary studies.

References

Anderson, Cormac, Johann-Mattis List & Tiago Tresoldi. 2018. Modelling sound change with the help of multi-tiered sequence representations. Paper presented at the 48th Poznań Linguistic Meeting, Poznań, 15 September 2018. https://bit.ly/3NtRK7F (accessed 10 January 2020).

Baker, Adam. 2008. Computational approaches to the study of language change. *Language and Linguistics Compass* 2. 289–307.

Barnett, Phillip & Ryan Macdonald. 2018. DERBiPIE: A computational database for Proto-Indo-European language. Paper presented at the Thirtieth Annual UCLA Indo-European Conference, November 9–10, 2018, University of California, Los Angeles.

Bergsland, Knut & Hans Vogt. 1962. On the validity of glottochronology. *Current Anthropology* 3. 115–53.

Bouckaert, Remco, et al. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337. 957–60.

Burton-Hunter, Sarah K. 1976. Romance etymology: A computerized model. *Computers and the Humanities* 10. 217–20.

Campbell, Lyle. 2013. *Historical linguistics: An introduction*. 3rd ed. Edinburgh: Edinburgh University Press.

Chang, Will, Chundra Cathcart, David Hall & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91. 194–244.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York: Harper & Row.

Chrétien, Charles Douglas & Alfred Kroeber. 1937. Quantitative classification of Indo-European languages. *Language* 13. 83–103.

Clackson, James. 2007. *Indo-European linguistics: An introduction*. Cambridge: Cambridge University Press.

Cowgill, Warren. 1966. A search for universals in Indo-European diachronic morphology. In Joseph Greenberg (ed.), *Universals of language*. 2nd ed., 114–41. Cambridge, MA: MIT Press.

Czekanowski, Jan. 1927. *Wstęp do historji Słowian [Introduction to the history of the Slavs]*. Lwów: Jakubowski.

Dunn, Michael. 2015. Language phylogenies. In Claire Bowern & Bethwyn Evans (eds.), *Routledge handbook of historical linguistics*, 190–211. Abingdon: Routledge.

Dyen, Isidore, Joseph Kruskal & Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82. 1–132.

Eastlack, Charles L. 1977. Iberochange: A program to simulate systematic sound change in Ibero-Romance. *Computers and the Humanities*, 11. 81–8.

Eichner, Heiner. 1992. Indogermanisches Phonemsystem und lateinische Lautgeschichte. In Oswald Panagl & Thomas Krisch (eds.), *Latein und Indogermanisch*, 55–79. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck.

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

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

Goddard, Ives. 1998. Recovering Arapaho etymologies by reconstructing forwards. In H. Craig Melchert & Jay H. Jasanoff (eds.), *Mír curad: Studies in honor of Calvert Watkins*, 183–200. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck.

Gray, Russell & Quentin Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426. 435–9.

Greenhill, Simon. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics* 37. 689–98.

Hartman, Steven Lee. 2003. Phono (Version 4.0): Software for modeling regular historical sound change. In Leonel Ruiz Miyares, Celia E. Álvarez Moreno & María Rosa Álvarez Silva (eds.), *Actas: VIII Simposio Internacional de Comunicación Social*. Vol. 1, 606–9. Málaga: Universidad de Málaga & Santiago de Cuba: Centro de Lingüística Aplicada.

Hartman, Steven Lee. 2018. Phono, software for historical phonology. https://langnhist.weebly.com/phonoTOC.html (accessed 10 January 2020).

Hewson, John. 1974. Comparative reconstruction on the computer. In John M. Anderson & Charles Jones (eds.), *Historical linguistics.* Vol. 1. *Syntax, morphology, internal and comparative reconstruction*, 191–7. Amsterdam: North-Holland.

Hock, Hans. 2012. Phrasal prosody and the Indo-European verb. In H. Craig Melchert (ed.), *The Indo-European verb*, 115–26. Wiesbaden: Reichert.

Hoijer, Harry. 1954. Lexicostatistics: A critique. *Language* 32. 49–60.

Jäger, Gerhard. 2018. Computational historical linguistics. ArXiv:1805.08099. https://arxiv.org/abs/1805.08099 (accessed 10 January 2020).

Janhunen, Juha & Jouna Pyysalo. 2018. Indo-European linguistics in the 21st century (1): From trilaryngealism to monolaryngealism – returning to Oswald Szemerényi. *Urindogermanische Sprachforschungen* 2018(1). 1–14.

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

Kessler, Brett. 2004. Derive. http://spell.psychology.wustl.edu/derive/ (accessed 10 January 2020).

Kim, Ronald I. 2018. Greco-Armenian: The persistence of a myth. *Indogermanische Forschungen* 123. 247–72.

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

Kortlandt, Frederik. 2007. *Italo-Celtic origins and prehistoric development of the Irish language*. Amsterdam: Rodopi.

Kortlandt, Frederik. 2008. Balto-Slavic phonological developments. *Baltistica* 43. 5–15.

Liesner, Malte. 2015. *Greek historical phonology workbook*. Wiesbaden: Reichert.

List, Johann-Mattis. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the EACL 2017 Software Demonstrations*, 9–12. Stroudsburg, PA: Association for Computational Linguistics.

Lubotsky, Alexander M. 2018. The phonology of Proto-Indo-Iranian. In Jared S. Klein, Brian D. Joseph & Matthias Fritz (eds.), *Handbook of comparative and historical Indo-European linguistics*. Vol. 3, 1875–88. Berlin: De Gruyter Mouton.

McCone, Kim. 1996. *Towards a relative chronology of ancient and medieval Celtic sound change*. Maynooth: Department of Old Irish, St. Patrick's College.

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

McMahon, April & Robert McMahon. 2006. *Language classification by numbers*. Oxford: Oxford University Press.

Maniet, Albert. 1985. Un programme de phonologie diachronique: De l'"indoeuropéen" au latin par ordinateur – version définitive. *Cahiers de l'Institut de Linguistique de Louvain* 11. 203–43.

Matasović, Ranko. 2005. Toward a relative chronology of the earliest Baltic and Slavic sound changes. *Baltistica* 40. 147–57.

Meillet, Antoine. 1922. *Les dialectes indo-européens*. 2nd ed. Paris: Champion.

Meiser, Gerhard. 2003. *Veni, vidi, vici: Die Vorgeschichte der lateinischen Perfektsystems*. Munich: Beck.

Mittmann, Roland. 2009. Ein Verfahren zur Ermittlung der relativen Chronologie der vorgotischen Lautgesetze. In Christian Chiarcos, Richard Eckart de Castillo & Manfred Stede (eds.), *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From form to meaning: Processing texts automatically*, 199–209. Tübingen: Narr.

bin Muzzafar, Towhid. 1997. *Computer simulation of Shawnee historical phonology*. Unpublished MA thesis, Memorial University of Newfoundland.

Nakhleh, Luay, Don Ringe & Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81. 382–420.

Nichols, Johanna & Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2. 760–820.

Noyer, Rolf. 2016. *Etymological database for Indo-European languages*. https://bit.ly/3wJU5EK (accessed 10 January 2020).

Olander, Thomas. 2015. *Proto-Slavic inflectional morphology: A comparative handbook*. Leiden: Brill.

Pereltsvaig, Asya & Martin W. Lewis. 2017. *The Indo-European controversy: Facts and fallacies*. Cambridge: Cambridge University Press.

Pyysalo, Jouna. 2017. Proto-Indo-European lexicon: The generative etymological dictionary of Indo-European languages. In Jörg Tiedemann & Nina Tahmasebi (eds.), *Proceedings of the 21st Nordic Conference of Computational Linguistics*, 259–62. Linköping: Linköping University Electronic Press.

Renfrew, Colin. 1987. *Archaeology and language: The puzzle of Indo-European origins*. London: Penguin.

Ringe, Don, Tandy Warnow & Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100. 59–129.

Ringe, Don. 2012. Cladistic principles and linguistic reality: The case of West Germanic. In Philomen Probert & Andreas Willi (eds.), *Laws and rules in Indo-European*, 33–42. Oxford: Oxford University Press.

Ringe, Don. 2017a. *From Proto-Indo-European to Proto-Germanic*. 2nd ed. Oxford: Oxford University Press.

Ringe, Don. 2017b. Indo-European dialectology. In Jared S. Klein, Brian D. Joseph & Matthias Fritz (eds.), *Handbook of comparative and historical Indo-European linguistics*. Vol. 1, 62–75. Berlin: De Gruyter Mouton.

Safarewicz, Jan. 1948. Krytyka metody ilościowej stosowanej w ocenie pokrewieństwa językowego [A critique of the quantitative method used in evaluating linguistic relationship]. *Biuletyn Polskiego Towarzystwa Językoznawczego* 8. 30–9.

Schmuki, Marcel. 2001. ETYMO. http://www.etymo.net/ (accessed 10 January 2020).

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press

Schrijver, Peter. 2006. Review of Meiser 2003. *Kratylos* 51. 46–64.

Skousen, Royal. 1989. *Analogical modeling of language*. Dordrecht: Kluwer.

Sims-Williams P. 2018. Mechanising historical phonology. *Transactions of the Philological Society* 116. 555–73.

Smith, Raoul N. 1969. A computer simulation of phonological change. *ITL: Tijdschrift voor Toegepaste Linguistiek* 5. 82–91.

van der Staaij, Robert. 1995. *A reconstruction of Proto-Italic*. Unpublished PhD thesis, Leiden University.

Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96. 452–63.

Szemerényi, Oswald. 1996. *Introduction to Indo-European linguistics*. Oxford: Oxford University Press.

Tischler, Johann. 1973. *Glottochronologie und Lexikostatistik*. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck.

de Vaan, Michiel. 2008. *Etymological dictionary of Latin and the other Italic languages*. Leiden: Brill.

Watkins, Calvert. 1962. *Indo-European origins of the Celtic verb*. Dublin: Dublin Institute for Advanced Studies.

Weiss, Michael. 2020. *Outline of the historical and comparative grammar of Latin*. 2nd ed. Ann Arbor, MI: Beech Stave.

Wichmann, Søren, Eric Holman & Cecil Brown (eds.). 2018. The ASJP database (version 18). http://asjp.clld.org/.

Wodtko, Dagmar, Britta Irslinger & Carolin Schneider (eds.). 2008. *Nomina im indogermanischen Lexikon*. Heidelberg: Winter.

Zair, Nicholas. 2018. The shared features of Italic and Celtic. In Jared S. Klein, Brian D. Joseph & Matthias Fritz (eds.). 2018. *Handbook of comparative and historical Indo-European linguistics*. Vol. 3, 2030–7. Berlin: De Gruyter Mouton.

Zeldes, Amir. 2008–9. Sound change transducers. https://corpling.uis.georgetown.edu/amir/sct.php (accessed 10 January 2020 but no longer available).

https://doi.org/10.1017/9781108758666.003 Published online by Cambridge University Press