

# Subgrouping Indo-European: A fresh perspective\*

Robert Mailhammer (Arizona State University)

Robert.Mailhammer@asu.edu

## 1. Introduction

For several reasons the Indo-European languages present significant challenges to subgrouping. First, the long research history is an advantage in terms of data and research, but it also represents a complication resulting from diverging avenues of research and research traditions. Second, while the long history of attestation is beneficial for historical research, because of the time depth of available data, the unbalanced distribution of attestation and the fact that the time depth of attestation is frequently still insufficient pose particular problems for subgrouping. And third, there are special issues in subgrouping Indo-European which make it doubtful whether the family has clean binary splits above branch-level. Rather, it seems to be the multilateral affinities among the Indo-European languages and branches which point to a different genesis. As a result, there appears to be a bit of a deadlock in this area, which may well be broken by new approaches, such as the Automated Similarity Judgement Program (ASJP; <http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm>).

This contribution compares the way the Indo-European languages are subgrouped by the ASJP algorithm, which is based on the phonetic similarity of lexical items, to the consensus classification according to the Comparative Method. It will show that despite its fairly crude method ASJP is actually reasonably accurate in identifying branches. However, there are problems below branch-level, in particular with respect to the identification of sub-branches. Additional complications result from the nature of the data. This concerns both the accuracy of the data and the fact that any lexicostatistic method will, of course, be affected by high proportions in loanwords if they are not detected (loanwords are generally removed from the datasets). This is evident from old loanwords, such as E *person* and *mountain*, which can differ significantly from inherited words, cf. e.g. G *Mensch, Berg*, Du *mens, berg*, and so

---

\*This is the written up and revised version of a talk given at the ALP conference in Leipzig, 18 September 2010. Thanks to the organisers for inviting me and the audience for their feedback; thanks also everyone at the MPI for Evolutionary Anthropology for their support. I am also grateful to Peter-Arnold Mumm (Munich) for valuable suggestions and helpful information on certain subgroups of Indo-European, most notable Iranian.

forth. Fortunately, the number of loanwords among the 40 items used by ASJP tends to be low, but in extreme cases this may well be a decisive factor. For instance, the lexicon of Albanian contains a large number of Latin loanwords, a fact that is well-known and that could have affected the grouping according to the ASJP method (see fn 3 below). I will also show that the ASJP tree picks up strands of older research, such as the Italo-Celtic hypothesis and the long-assumed Greco-Armenian subgroup. Finally, I will comment on the different results the ASJP method yields for the subgrouping of the Indo-European languages with relation to the size and nature of the database, comparing the Indo-European ASJP tree with the position of the Indo-European languages on the ASJP world tree.

The structure of this paper is as follows. Section 2 gives an overview of the Indo-European language family and its branches, according to the general opinion. In section 3 discuss some problems and topics in subgrouping Indo-European, while section 4 compares the traditional subgroups of Indo-European arrived at by virtue of the Comparative Method to the trees of the ASJP project. In particular I will deal with the following questions:

- How well do the trees match?
- Does the ASJP tree reflect facts about the prehistory of the Indo-European languages?
- Does the ASJP tree pick up on geographical proximity/distance?
- What are the implications of the ASJP world tree for subgrouping Indo-European?

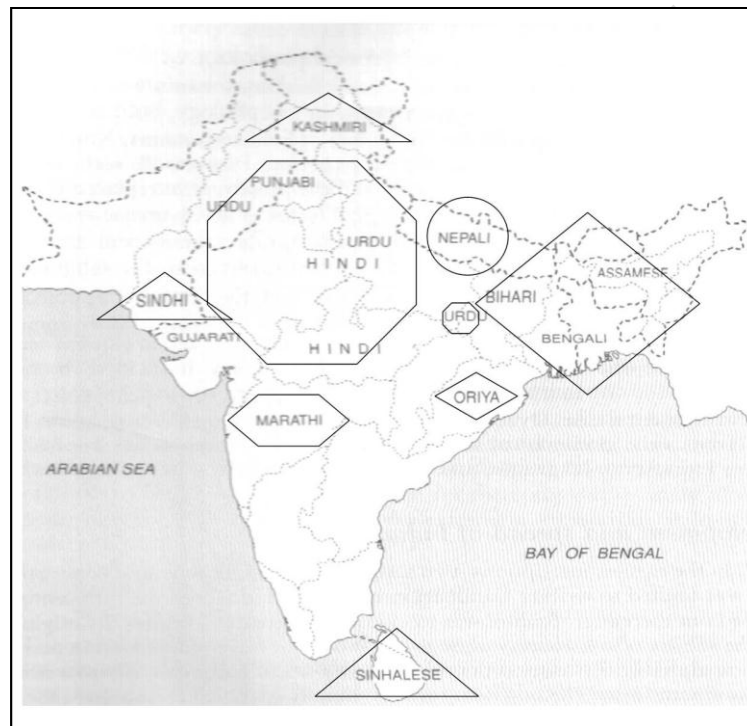
The final section wraps up with a summary and some concluding remarks.

## **2. The Indo-European language family**

The Indo-European languages form a large language family with a long documented history and a wide geographical distribution (see Map 1 below). In crosslinguistic comparison, they appear typologically similar, but they do in fact exhibit considerably diversity with respect to certain parameters, e.g. word order. In addition, some branches with high significance to reconstructing the Indo-European protolanguage have become extinct, most notably Anatolian and Tocharian. Furthermore, while some branches show a fair amount of high dialectal diversification, e.g. Germanic, others appear to represent essentially the same language, e.g. Greek.



Modern Indic languages are spoken by a fifth to a sixth of today's population, and show a number of features that are believed to be the effect of a Dravidian substrate (e.g. the retroflex consonants of Hindi), but also typologically unusual phenomena, which are the result of internal developments, such as split ergativity (also found in Iranian and transparently going back to a construction in Sanskrit, see e.g. Fortson 2010: 221). Map 2 depicts a selection of key Modern Indic languages as well as their subdifferentiation.



**Map 2:** Modern Indic languages (adapted from Fortson 2010: 222)

According to the general opinion (Fortson 2010: 222–223), the Indic languages are grouped into an eastern zone (diamond areas in Map 2), a northern zone (circle), a central zone (octagons), a southern zone (hexagon) and a northwestern zone (triangles). Note in particular the geographical split of the northwestern zone comprising Kashmiri in the far north, Sindh in the west and Sinhalese in the far south. Furthermore, it is significant with respect to the ASJP that Urdu and Hindi are structurally essentially the same language, though showing considerable differences in their lexicons, with Urdu possessing a large amount of Arabic loanwords.

## 2.2 Iranian

The Iranian languages are spoken over a wide geographical area. Similar to their Indic relative, Iranian languages have also developed split ergativity through an internal

development. As far as subdifferentiation is concerned, commonly an eastern and a western branch are recognised. Map 3 contains important modern Iranian languages and their classification (East Iranian underlined in red, West Iranian in blue).



**Map 3:** Important modern Iranian languages (adapted from Fortson 2010: 244)

### 2.3 Greek

Greek is among the Indo-European languages with the oldest attestations. However, it is significant that there is an attestational gap of about 600 years, which is often referred to as the “Dark Ages”. The oldest records are only from one dialect, Mycenaean, whereas the later records render a picture of a considerable dialectal differentiation but no subdifferentiation as far as sub-branches, i.e. languages, are concerned. It has often been remarked that Greek superseded earlier languages, possibly both non-Indo-European and Indo-European languages (see Strunk 2003 and Garrett 2006), and also that Greek contains a substantial number of non-Indo-European loanwords and non-etymologised vocabulary (see e.g. Mailhammer 2007 for an etymological quantification of the Greek primary verbs).<sup>1</sup>

### 2.4 Italic

Although the Italic contains other branches than Latin, all surviving languages are descendants of Latin (the Romance languages). This puts the modern representatives of this

<sup>1</sup> See also Tischler 1979: 267 for a discussion of Greek in relation to Hittite.

branch in the comfortable position of having a richly attested common ancestral language, a situation they do not share with many Indo-European subgroups. With the exception of Romanian, all modern Romance languages share significant parts of their vocabularies. Though all have a fair amount of loanwords, this generally does not extend to the relevant part of the core vocabulary. Moreover, modern Romance languages tend to be similar with respect to their sound inventories, especially in terms of consonants, which can be expected to have an impact on ASJP classification.

## 2.5 Celtic

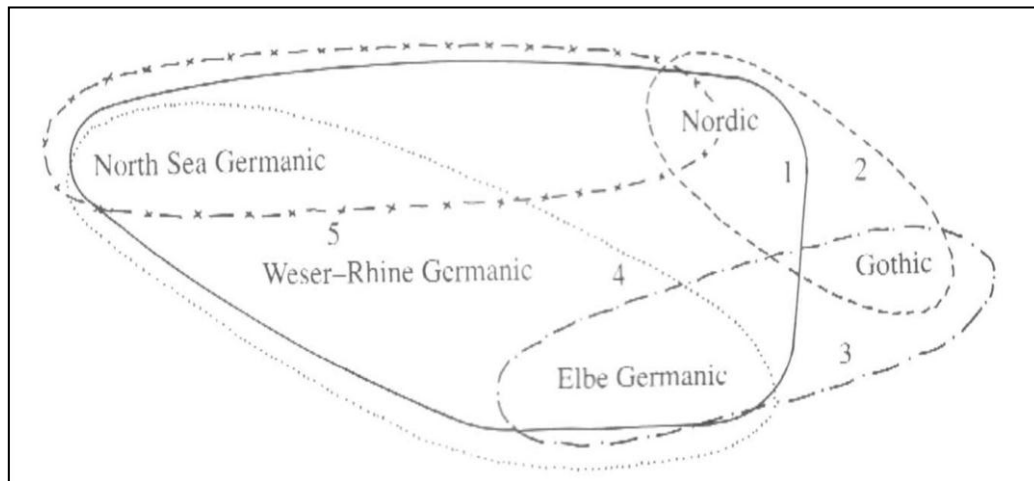
The Celtic branch is commonly divided into two subgroups, Continental and Insular Celtic, of which the former is extinct and so poorly attested that its documentation can only be called insufficient.<sup>2</sup> By contrast, both groups of Insular Celtic, Brittonic and Goidelic are well-attested, documented and researched. Modern representatives of Insular Celtic are Welsh, Breton, Manx (Brittonic) and Scottish and Irish Gaelic (Goidelic).

## 2.6 Germanic

Like most Indo-European branches, the protolanguage of Germanic is only indirectly attested via its daughter languages. Between Proto-Germanic and the earliest attestations of a Germanic daughter language (Runic Norse) there is a gap of several hundred years. Nevertheless, Proto-Germanic is relatively well-reconstructed, but there are some issues with the subgroups of Germanic which are symptomatic for subgrouping the Indo-European languages as a whole. The commonly accepted division is between North, West and East Germanic. All East Germanic languages, e.g. Gothic and Burgundian, have become extinct, so modern Germanic languages are either North or West Germanic. North Germanic comprises basically all Scandinavian languages. West Germanic is commonly subdivided into North Sea Germanic (English, Frisian), Low German (Old Saxon, Old Low Franconian and their main descendants Low German and Dutch) as well as High German (descendants of Old High German and Lombardic). However, this traditional classification, though well justified on several grounds, is criss-crossed by isoglosses running orthogonal to it, which becomes clear from the following diagram taken from Ramat (2006).

---

<sup>2</sup> There is an alternative classification according to the reflex of Proto-Celtic *\*kw-*, differentiating P-Celtic (Brittonic and Gaulish,) from Q-Celtic (e.g. Schmidt (1988)). Because it is based only on a single feature, it has not found common acceptance among the majority of Celticists (see MacCone (1996) for a discussion).



**Figure 1:** Dialectal affinities among the Germanic languages (Ramat 2006: 385)

One example for such an orthogonal isogloss is the reflex of the Proto-Germanic voiceless stops in the Germanic daughter languages. While all other dialects do not change them, High German and Lombardic develops them to affricates and fricatives, thus creating an opposition of what Vennemann (1984*et passim*) calls Low vs. High Germanic. Despite this heterogeneous picture, the traditional classification mentioned above represents the general opinion in Germanic scholarship and has not seriously been challenged, so that it is used as reference point here.

## 2.7 Slavic

Although there is solid indication that the Slavic and Baltic languages have descended from a common node, this section focuses only on Slavic, since issues of groupings above branch level will be discussed in section 3 below. Traditionally, the Slavic languages are subdivided into East (Russian, Belorussian and Ukrainian), West (Polish, Czech, Sorbian and Slovak) and South Slavic (Slovenian, Serbo-Croatian, Macedonian and Bulgarian). The oldest attested Slavic language is Old Church Slavonic (from the second half of the 9<sup>th</sup> century AD), which basically belongs to the southern branch, but shows affinities to all branches, depending on where a given text was composed. According to the general opinion, Old Church Slavonic comes close to the reconstructed ancestor of the Slavic languages, Proto-Slavic (see e.g. Tichy 2004: 15).

## 2.8 Baltic

The Baltic languages form a small branch comprising Lithuanian, Latvian (East Baltic) and the extinct Old Prussian (West Baltic). As mentioned above, it is commonly thought that Baltic and Slavic together descended from a Proto-Balto-Slavic ancestor. Furthermore,

Lithuanian displays some remarkably ancient traits, particularly with respect to its phonology. As in other branches, the Baltic protolanguage is not directly attested, the earliest Baltic attestations (in Old Prussian) date from the 14<sup>th</sup> century AD (Meier-Brügger 2002).

## 2.9 Armenian

Until the publication of Hübschmann (1877) Armenian was believed to be an Iranian dialect. This is because of the high number of loanwords from Iranian (but also Greek and Syriac), especially when set against the only 450 attested genuine Armenian words (Fortson 2010: 383; see Mallory & Adams 2006 for a list of the native Armenian words). But Hübschmann pointed out significant phonological and morphological characteristics that divided Armenian and Iranian, establishing Armenian as a branch in its own right. Armenian is first attested in the middle of the 5<sup>th</sup> century AD, and interestingly there are no attestations warranting a further dialectal subdivision.

## 2.10 Albanian

Albanian is attested from the 15<sup>th</sup> century AD onwards. It constitutes a branch though affiliations have been debated. Albanian contains a huge amount of Latin loanwords, and is divided into two dialects, northern Gegic and southern Tosk.

## 2.11 Extinct branches of Indo-European

Besides sparsely attested languages, such as Phrygian, there are two extinct branches of Indo-European which are of high significance with respect to subgrouping the language family and to reconstructing the Indo-European protolanguage, namely Anatolian and Tocharian. However, since the ASJP tree is based on living languages, these two branches will be dealt with only briefly here (see Appendix B for a discussion of an ASJP tree based on earliest attested stages of the Indo-European languages). Anatolian boasts the oldest attestations of Indo-European, dating from the 16<sup>th</sup> century BC, the chief witness being Hittite, and it shows striking peculiarities in comparison with other Indo-European language. Tocharian is attested from the 6<sup>th</sup> to the 8<sup>th</sup> century AD in two versions, called Tocharian A and B. Both branches have contributed significantly to the reconstruction of Proto-Indo-European, but Anatolian has posed the greater riddles for reconstruction and subgrouping. Despite clearly conservative features in phonology (laryngeals) and morphology (e.g. only two genders), Anatolian lacks a number of properties that are usually posited for Proto-Indo-European, such as the dual and a number of TAM categories. Consequently, it is disputed whether Anatolian is a sister to Proto-Indo-European in terms of subgrouping or whether it split off from the common

protolanguage at some point. With respect to the latter alternative there is debate whether Hittite transformed certain categories it lacks, such as the aorist, or whether it split off so early that it never developed them. The consensus view is that Anatolian is a descendant of Proto-Indo-European (like all other branches), and that it does not modify the general reconstruction of Proto-Indo-European based mainly on Indo-Iranian and Greek. Other extinct Indo-European languages, such as Phrygian and Venetian, are not attested richly enough to make reliable statements about subgrouping.

### 3. Subgrouping Indo-European

It is a striking feature of the Indo-European language family that above branch-level there are only few clear-cut groups that can reliably be put together. The research history on Indo-European languages documents the various attempts to form higher-node subgroups, and shows that failed attempts have by and large led to a more sober approach (a good overview of the problem summarising traditional research is Porzig 1954). However, as will become clear below, in its foundations, the modern discussion on Indo-European subgrouping essentially revolves around the same issues that made Johann Schmidt and Hugo Schuchardt argue for a modification of Schleicher's family tree model in the late 19<sup>th</sup> century. What it boils down to is a controversy between "lumpers" and the "splitters", and this is not a debate that is confined to Indo-European. But what makes this particular problem so difficult to solve in a neat way is that there is a tremendous amount of data and a long research history. The former means that all but a few propositions of subgroups face contradicting isoglosses arguing for a different classification. The consequence of the latter point is that the research community is somewhat fragmented, synchronically as well as diachronically, as there are many people and opinion that may be in competition. For instance, while there is a traditional research strand connecting Armenian with Greek in one subgroup, there is probably good indication that this is due to contact rather than close genetic relation (see below).

In a nutshell, the modern discussion on subgrouping Indo-European centres on the question whether Proto-Indo-European split up in a rake-like fashion with only few branches (if any) being joined up at a higher level (see Figure 2), or whether Proto-Indo-European split up in a largely binary fashion (see Figure 3).

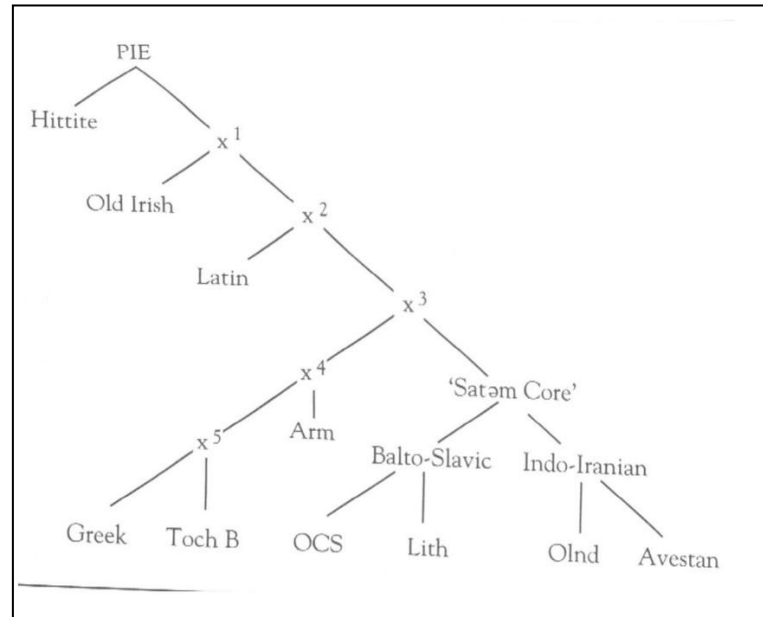


Figure 2: Modified rake model of IE (Mallory & Adams 2006: 80)

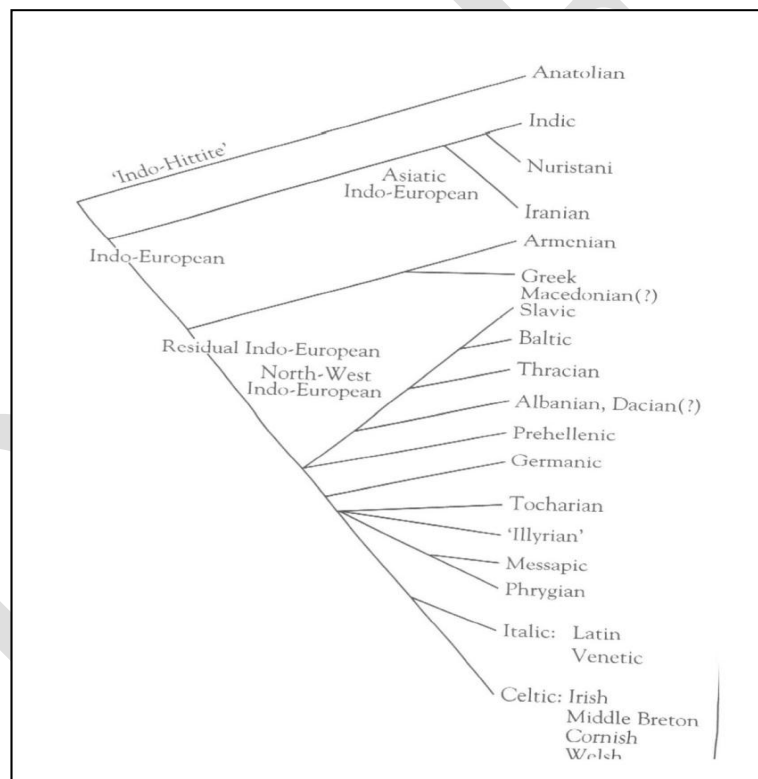


Figure 3: A binary tree of Indo-European (Mallory & Adams 2006: 74)

As mentioned above, most attempts at forming higher-order subgroups have been met with strong criticism. Two prominent examples are Armenian-Greek and Italo-Celtic. After reviewing the evidence, Clackson (1994) draws the sobering conclusion that the similarities between the Greek and Armenian do not warrant the assumption of a closer genetic connection but that they are the result of contact. Likewise, Sims-Williams (2006) argues

convincingly against the Italo-Celtic hypothesis (see also Meier-Brügger 2002) with references. The two exceptions are Indo-Iranian and Balto-Slavic, both of which are commonly accepted subgroups of Indo-European.

However, there is also a methodological dimension to the subgrouping debate in Indo-European (and this is also reflected in similar discussions for other families, e.g. Australian languages). This concerns the question whether subgroups are best based on lexical material or on structural features or both. While wordlists and respective etymologies are both the historical tool – Hübschmann’s arguments arose from etymologies – as well as an indispensable tool in data-impooverished areas such as Indigenous Australia, many would argue that structural features – in particular, morphology – should prove more reliable. In fact, given the likelihood of parallel innovation in phonology and syntax, many, such as Ringe et al. (2002) have argued in favour of a more comprehensive set of data. If that is accepted, however, bearing in mind that, methodologically, subgroups are identified by shared innovations rather than by shared retentions, one pertinent question is how many features are needed to convincingly argue for a sub group. That this is not an academic question is proved by the numerous attempts to classify Indo-European into higher-order groups based on various criteria. For instance, the traditional *centum-satem* distinction worked fine until Hittite and Tocharian were discovered. Another famous differentiation involves the first sound of the dative/instrumental endings, *\*-m-* vs. *\*-b<sup>h</sup>-*, which Watkins (2006: 30) casts serious doubt upon by showing that this dichotomy may be explained without proposing different subgroups. Even an elaborate method using a mixed set of lexical and structural data like that of Ringe et al. (2002) runs into problems stemming from the interdialectal affinities of some Indo-European languages, in their case, Germanic (see Ringe et al. 2002: § 7). Consequently, while in the case of Indo-European one feature certainly does not seem to be enough, it is far from clear that more features bring better result. Figure 4 shows some of the affinities between the branches of Indo-European.

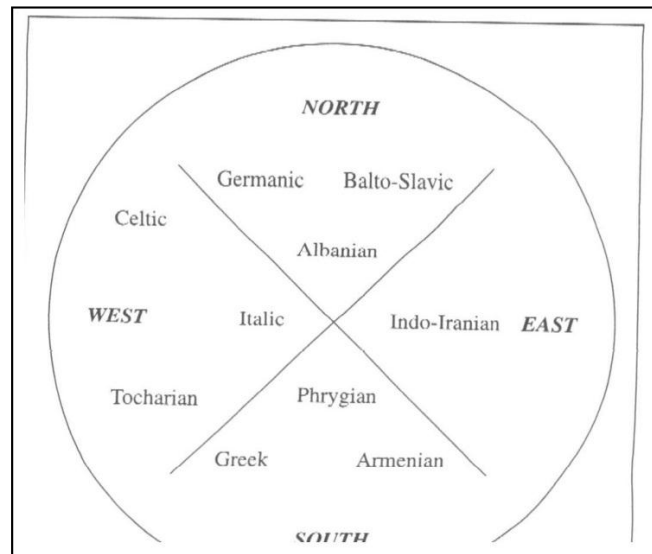


Figure 4: A diagram of dialectal affinities (Watkins 2006: 30)

The situation is further complicated by other factors, such as potential feature diffusion (see already Porzig 1954) as well as lost contact languages, both Indo-European and non-Indo-European, forming different strata (see Mailhammer *forthc.*). Taking such complexities into account, Garrett (2006) implicitly argues for the minimum consensus, which is the rake-model:

If this framework is appropriate for IE branches generally, we cannot regard IE ‘subgroups’ as subgroups in a classical sense. Rather, the loss or pruning’ of intermediate dialects, together with convergence *in situ* among the dialects that were to become Greek, Italic, Celtic, and so on, have in tandem created the appearance of a tree with discrete branches. But the true historical filiation of the IE family is unknown, and it may be unknowable.

To sum up, according to the current knowledge, the modified rake-model with Indo-Iranian and Balto-Slavic as the only clear nodes above branch-level (cf. Figure 2 above) is taken as the specialist subgrouping of Indo-European the ASJP classification is set against.

#### 4. The Comparative Method vs. The Automated Similarity Judgement Program

In order to examine how well the ASJP classification lines up with the conventional subgrouping of Indo-European, it is useful to recall some fundamental differences between the two methods. The point of departure from which the Comparative Method (CM) sets out is always the oldest attestation, which, in this case, means that it mandatorily involves extinct languages. This is different for the main classification by the ASJP dealt with here, as it is based only on living Indo-European languages (as mentioned above, an ASJP version based on older stages of Indo-European languages including Hittite and Tocharian is briefly

discussed in Appendix B). Another methodological difference is that the CM takes into account lexical as well as structural correspondences, whereas by virtue of method, the ASJP only considers lexical data. Furthermore, the CM usually is conducted by human researchers, but this is precisely the factor the ASJP aims to eliminate from the equation. Finally, the ASJP method does not work with actual cognates but on the basis of phonetic similarity as calculated by an algorithm with only an implicit understanding that closed phonetic similarity is likely due to a genetic relation. Consequently, the CM clearly is the more thorough method, in particular, as generations of researchers have investigated Indo-European. However, given the time depth of the Indo-European languages and the crudeness of the method, this section will show that it does rather well. For reasons of space the whole ASJP tree can be found in Appendix A. The following sections focus on specific problems.

#### 4.1 Issues at branch-level

Although ASJP picks up the two consensus nodes Balto-Slavic and Indo-Iranian as well as Romance, Greek and Germanic as branching off from Proto-Indo-European, the positions of Celtic, Albanian and Armenian are different to that of the consensus grouping according to the comparative method.

In the ASJP classification Albanian is grouped with the Romance languages, against the consensus view, which usually sees Albanian as an independent branch of Indo-European. This may be due to the large number of Latin loanwords in Albanian, even more so in the Tosk variety, which is spoken in southern Albania and also in Italy.<sup>3</sup>

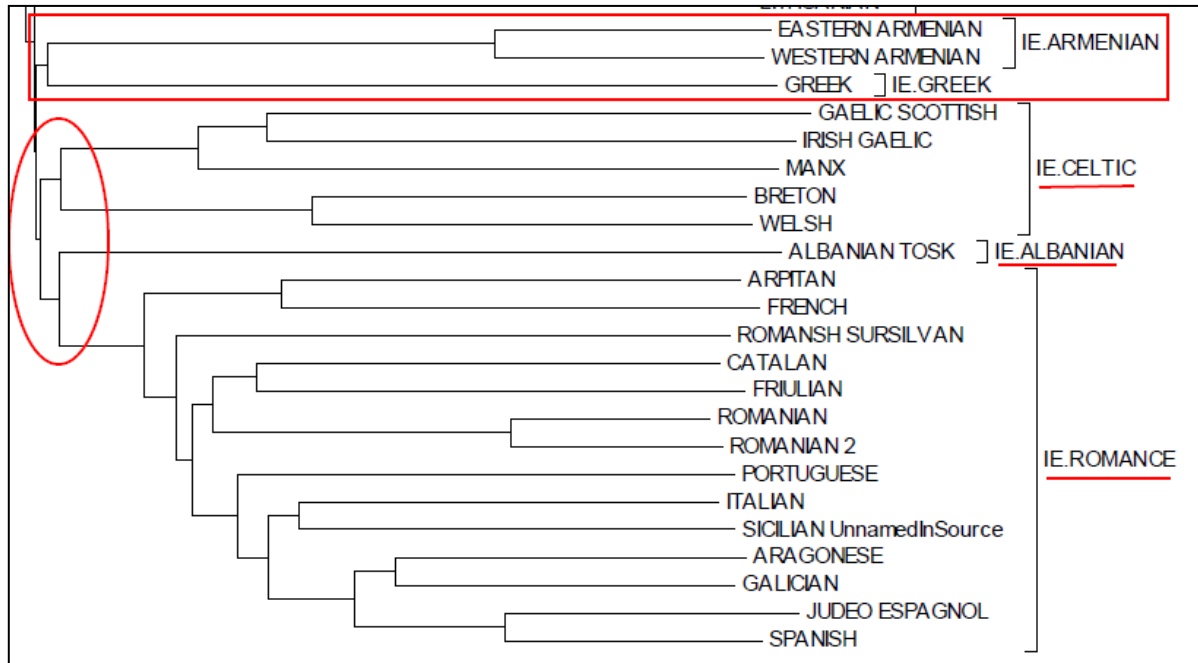
Another difference between the consensus classification and ASJP is that Celtic belongs to a branch that also includes Albanian and the Romance languages. As pointed out above, the traditionally assumed Italo-Celtic node cannot be sufficiently substantiated (see 3. above).

And finally, the ASJP tree diverges from the current Indo-European consensus classification in grouping Armenian and Greek together under one node. As mentioned in section 3 above, the similarities of both languages are more likely due to contact rather than

---

<sup>3</sup> For example, the Albanian word translating ‘come’ (*vij*) in the ASJP database is possibly a Latin loanword, the native word derived from the Proto-Indo-European root means ‘run, haste’ Demiraj . Of course, this would have to be verified independently by an etymology for *vij*. In addition, ‘tongue’, ‘person’ and ‘sun’ have no entries in Demiraj’s *The online database of the Albanian inherited lexicon* Demiraj , which also makes it likely that these words are actually loanwords. To be sure, a comprehensive investigation would have to be undertaken to assess the proportion of non-Albanian words in the ASJP dataset.

to a closer genetic relationship. Figure 5 shows the relevant section of the ASJP tree for Indo-European.

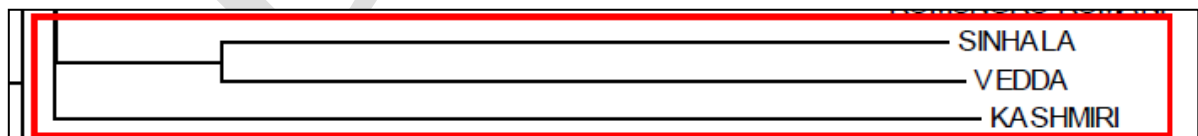


**Figure 5:** Section of ASJP tree showing the grouping of Armenian, Greek, Celtic, Albanian and Romance

## 4.2 Subgrouping within branches

### 4.2.1 Indic

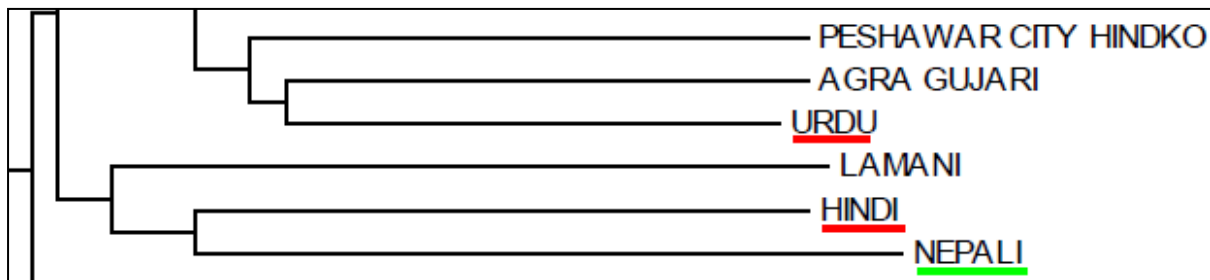
First of all, it is noteworthy that Panjabi is missing from the ASJP tree, though this is really a minor point, as this is clearly attributable to a lack of data. Second, the ASJP classification correctly identifies the genetic relation between Kashmiri spoken in the north and Sinhalese in the south, showing that geographic separation in this case has not impacted on the 40 lexical items used by ASJP (see Figure 6).



**Figure 6:** Northwestern Indic in the ASJP tree

Third, Urdu and Hindi, which are virtually the same language, are placed in different sub-branches by ASJP. This may be because of the large number of Arabic and Persian loanwords in Urdu, though this would have to be verified for the 40 words used for classification. In addition, Hindi is grouped with Nepali, which, according to the consensus classification,

forms its own branch within the Indic languages (see 2.1 above). The position of Hindi and Urdu on the ASJP tree can be seen in Figure 7.



**Figure 7:** The position of Urdu, Hindi and Nepali on the ASJP tree

#### 4.2.2 Iranian

The traditionally assumed division into an eastern and a western sub-branch does not come out clearly in the ASJP tree. Immediately below the top node there are two branches, namely Ossetian and everything else. This special position of Ossetic is mirrored in the traditional view of it being the “most western of the Eastern Iranian languages” (Peter-Arnold Mumm, Munich, e-mail dated 25 November 2010). However, the remaining group of Iranian languages does not break up into an eastern and a western group. Instead one sub-branch contains all West Iranian languages (except Belouchi, for which ASJP does not have data) plus eastern Wakhi, Shugni and Yaghnobi, whereas the other East Iranian languages are in the other sub-branch. This is all the more peculiar since Pashto effectively forms a linguistic island among West Iranian languages – which could well result in more similarity – as opposed to the languages that are lumped together with the West Iranian languages in the ASJP tree, all of which are spoken in the far east of the Iranian language area.

#### 4.2.3 Germanic

As mentioned in 2.6 above, the classification of the Germanic languages is not without its problems, but there is consensus about the main divisions and the subdivision of West Germanic. The ASJP tree shows several divergences, with respect to firstly the identification of sub-branches as well as secondly the sub-classification of the Low German sub-branch.<sup>4</sup>

<sup>4</sup> Interestingly, in the ASJP tree that is based on older Indo-European languages, Germanic comes out all right.

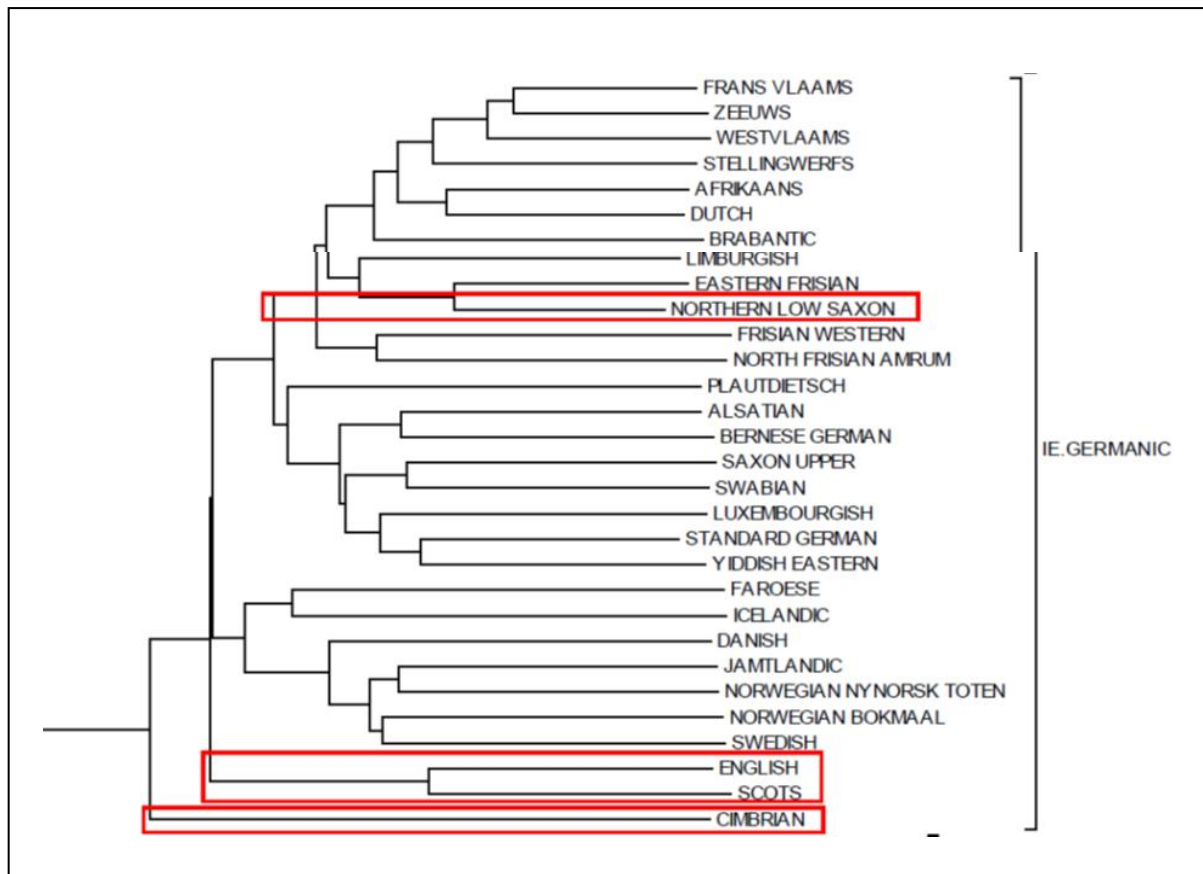


Figure 8: The ASJP tree for Germanic

On the topmost level there is the first divergence from the consensus classification, which is that Cimbric forms a branch of its own vs. all other Germanic languages, which are under a second node. As a Bavarian dialect, Cimbric should be grouped with the High German languages. In fact, the case of Cimbric illustrates a few problems with respect to data and coding the ASJP tree suffers from, which have also been reported for other families and languages. First of all, although Cimbric possesses a fair share of Italian loanwords, this is not an issue here, because they do not affect the basic vocabulary used by the ASJP data sources. But looking at the Cimbric source wordlist, there are cases which cause bewilderment. For instance, words like <zona> and <zeg> are highly doubtful, since the word-initial alveolar sibilant is generally voiceless in Bavarian. This makes a re-checking of the data necessary, since the coding in this case influences the calculation of phonetic similarity. A methodological problem is illustrated by cases of obviously inconsistent coding, which causes Cimbric to appear more different from its High German siblings than is actually the case: in the Standard German data source the diphthong [ɔʏ] is coded <oi>, as opposed to <oy> in Cimbric. It may be wondered whether the position of Cimbric changes once problems like these have been fixed.

Turning to the second node under which all remaining Germanic languages of the ASJP database are found, it can be noted that the North and West Germanic branches are identified correctly, but English and Scots form a separate branch (see Figure 8 above), which represents a clear difference to the consensus classification according to the Comparative Method. Accordingly, English belongs to the subgroup of West Germanic called North Sea Germanic, which also comprises Frisian.

While the North Germanic branch is fine, there are some issues with the West Germanic branch apart from the case of English/Scots just mentioned. Below the relevant node there are two distinct groups, one comprising High German dialects, and one comprising Dutch and its varieties, Frisian as well as Low German (cf. Figure 8 above). Even though High German as separate offshoot is in line with the consensus classification, the other part of this group is problematic in terms of genetic relationship.<sup>5</sup> Historically, Frisian belongs to the North Sea Germanic subgroup of West Germanic; one would therefore expect to have a distinct group branching off the West Germanic node comprising Frisian and English. Instead, West Germanic breaks up into what looks like a Low German group (plus Frisian) and a High German group (minus Cimbrian, see below). On a more fine-grained analysis, it is noteworthy that Northern Low Saxon is grouped together with Eastern Frisian, which may reflect geographic proximity, since it is clearly against the genetic classification. But then North and West Frisian, which are not spoken next to each other, share another intermediate node, and both groups together join up with Dutch and its other varieties (except for Limburgish) on the next level. Historically, Dutch and Low German belong to the same branch (deriving from the Low German dialects Old Low Franconian and Old Saxon), whereas Frisian does not. So this part of the Germanic tree clearly does not match the expert classification.

#### 4.3 Reflections of prehistory

Another question of interest is in what way the ASJP tree reflects the (pre-)history of the languages involved. While a detailed account would be a topic for a paper of its own, a few points may be mentioned. It is interesting that in the cases investigated here ASJP seems to pick up genetic similarity rather than geographical proximity though lexical similarity. For instance, despite the considerable geographic separation of Sinhala from its northern relative

---

<sup>5</sup> It is worth remembering that the ASJP method classifies according to lexical similarity, so the tree does not necessarily reflect genetic relatedness. I will return to this point further below.

ASJP correctly identifies both as members of one group. Similarly, in the Iranian branch, Pashto is correctly grouped with Eastern relatives rather than with West Iranian languages though it forms an island among these. Among the Germanic languages, Frisian and English are kept somewhat distinct among the West Germanic branches although a North Sea Germanic group as such is not identified, and although there are considerable problems with the subgrouping of West Germanic in the ASJP tree as a whole.

Apart from geographical proximity, by virtue of its method, ASJP can also identify more intimate contact between two languages, as this is likely to be reflected even in basic vocabulary. The clearest cases here are Armenian and Albanian, both of which have only very few native words, having borrowed extensively from Greek/Iranian and Latin respectively. Future research by specialists in Armenian and Albanian etymology may well reveal whether this is the cause for the position of both languages on the ASJP tree or whether this is due to some other reason.

#### 4.4 ASJP and older classifications

It is also interesting that the ASJP tree seems to pick up also higher-level groupings, which, although having been rejected in more recent literature, had been assumed in traditional research. This is valid for Italo-Celtic and Greco-Armenian, both of which have been shown to be superficial groupings (cf. 3. above). However, this is nevertheless remarkable, even more so as the ASJP tree for older Indo-European languages does not recognise any link between Italic and Celtic, but keeps that between Greek and Armenian (see Appendix B). In the Germanic branch ASJP misses North Sea Germanic, a long-recognised subgroup, while the two consensus groups, Balto-Slavic and Indo-Iranian, are identified without a problem.

This raises the question of whether ASJP could in fact uncover hitherto unrecognised connections. In the case of Greek and Armenian it clearly is a victim of its lexical basis, but is e.g. the lack of a clearly identifiable North Germanic group a mistake or a discovery? These questions have been raised in discussions within the ASJP project on a global scale, but a conclusion has not yet been reached.

#### 4.5 Indo-European in the ASJP world tree

Comparing the ASJP tree for Indo-European, i.e. a tree based only on Indo-European languages, and the ASJP world tree ([http://email.eva.mpg.de/~wichmann/language\\_tree.htm](http://email.eva.mpg.de/~wichmann/language_tree.htm)), i.e. a tree based on all living languages in the ASJP database, reveals a couple of interesting differences. First, the world tree contains also creoles based on Indo-European languages.

While this does not change the position of the Romance languages, it significantly changes the position of English and Scots, which are closely grouped with English-based creoles rather than with their other Germanic relatives. This in all likelihood is a result of the lexical method used by ASJP, since it is apparent that English-based creoles are essentially virtually identical with respect to their basic vocabulary, whereas languages like English and German, for instance, can exhibit considerable phonetic differences. Hence, this is clearly a method-based effect.

However, this does not suggest itself as an explanation for the second divergent case: The position of Greek on the world tree can only be described as “odd”. It is found in a subgroup together with the Nilo-Saharan (Koman) language Langa and members of different families further up the tree (see Figure 9).

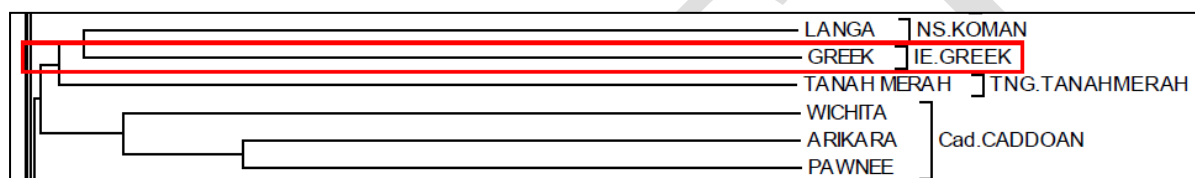


Figure 9: The position of Greek on the ASJP world tree

Finally, it is noteworthy that Celtic and Armenian form independent branches, thus agreeing with the consensus subgrouping according to the Comparative Method. While this seems like a good argument for basing ASJP trees on as much data as available, the two other cases, Greek and English, clearly argue against doing this. But it is significant that two datasets differing only in size have an influence on subgrouping according to lexical similarity. For English this is understandable since English-based creoles are virtual copies of English – just like Scots. But for Greek this is surprising, and it requires an explanation.

## 5. Conclusion

In summary, this paper compared the classification of Indo-European languages according to the ASJP methodology with the consensus subgrouping based on the Comparative Method. Although the ASJP method is in many ways a great deal cruder and although it was not designed for modelling genetic relationships, the ASJP tree matches the traditional classification rather well. It picks up all major branches of Indo-European, and it also correctly identifies the two higher node groupings, Indo-Iranian and Balto-Slavic. However, there are also some problems with respect to the position of languages and groups on the tree, e.g. Albanian (above branch level), Armenian and Greek (at branch level) and English

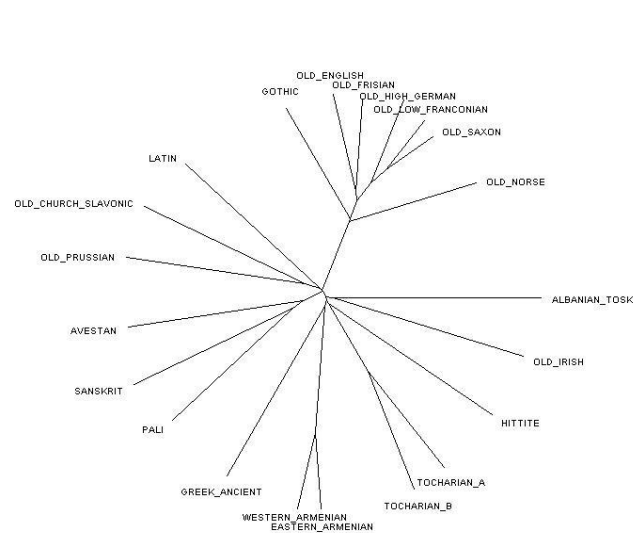
(languages). In addition, there are some data problems which influence subgrouping, as exemplified using Cimbrian. And finally, the methodology is sensitive to situations of intimate language contact with heavy borrowing, as seen in the cases of Urdu and Hindi, Albanian as well as Armenian.

As a result it can be concluded that ASJP methodology yields some good results for the classification of Indo-European languages, and that further refinement of the method is likely to yield even better results.

DRAFT



## Appendix B: ASJP tree based on older Indo-European languages



This tree differs in some respects from the ASJP tree in the main section of this paper, which is based on data from living languages only. First, Germanic is subgrouped correctly, clearly showing North, East and West Germanic, as well as the correct subdivision of West Germanic (North Sea Germanic, Low and High German). But Albanian is still in one branch with Celtic, though Latin has been separated, which raises interesting questions for the position of Albanian (but note that the Albanian data has not been changed to an older stage in contrast to all other languages, presumably because there was no older data available). Second, Armenian and Greek still form a subgroup, while nothing changes for Indo-Iranian and Balto-Slavic. Consequently, basing the tree for Indo-European on older languages shows some significant improvements, but cannot solve all problems identified in the main part of this paper. But it is a good argument for using older stages for the purpose of making inferences on genetic relationships within a language family.

## References

- Clackson, James. 1994. *The linguistic relationship between Armenian and Greek*. Univ. Diss.-Cambridge. (Publications of the Philological Society 30). Oxford: Blackwell.
- Demiraj, Bardhyl. The online database of the Albanian inherited lexicon. <http://www.indo-european.nl/index2.html> (25 November, 2010).
- Fortson, Benjamin W. 2010. *Indo-European language and culture: An introduction* (Blackwell textbooks in linguistics 19), 2nd edn. Chichester: Wiley-Blackwell.
- Garrett, Andrew. 2006. Convergence in the formation of Indo-European subgroups. Phylogeny and chronology. In Peter Forster & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 139–151. Cambridge: McDonald Institute for Archaeological Research.
- Hübschmann, Heinrich. 1877. *Armenische Grammatik*. Leipzig: Breitkopf & Härtel.
- MacCone, Kim. 1996. *Towards a relative chronology of ancient and medieval Celtic sound change* (Maynooth studies in Celtic linguistics 1). Maynooth: Dep. of Old Irish St. Patrick's College.
- Mailhammer, Robert. forthcoming. Diversity vs. uniformity: Europe before the arrival of the Indo-European languages: a comparison with prehistoric Australia. In Robert Mailhammer & Theo Vennemann (eds.), *Linguistic Roots of Europe*. Copenhagen: Museum Tusulanum Press.
- Mailhammer, Robert. 2007. *The Germanic strong verbs: Foundations and development of a new system*. Univ. Diss. u.d.T.: Mailhammer, Robert: A morphological and etymological study of the Germanic strong verbs--München, 2005. (Trends in linguistics Studies and monographs 183). Berlin: Mouton de Gruyter.
- Mallory, James & Douglas Q. Adams. 2006. *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European word*. Oxford: Oxford University Press.
- Meier-Brügger, Michael. 2002. *Indogermanische Sprachwissenschaft*, 8th edn. Berlin, New York: Walter de Gruyter.
- Porzig, Walter. 1954. *Die Gliederung des indogermanischen Sprachgebiets*. Heidelberg: Carl Winter.
- Ramat, Paolo. 2006. The Germanic Languages. In Anna G. Ramat & Paolo Ramat (eds.), *The Indo-European Languages*. London: Routledge.
- Ringe, Don. 2006. *From Proto-Indo-European to Proto-Germanic*. Oxford: Oxford University Press.
- Ringe, Don, Tandy Warnow & Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100. 59–129.
- Schmidt, Karl H. 1988. On the reconstruction of Proto-Celtic. In Gordon W. MacLennan (ed.), *Proceedings of the first North American Congress of Celtic Studies, held at Ottawa from 26th-30th March, 1986*, 231–248. Ottawa.
- Sims-Williams, Patrick. 2006. The Celtic Languages. In Anna G. Ramat & Paolo Ramat (eds.), *The Indo-European Languages*, 345–379. London: Routledge.
- Strunk, Klaus. 2003. 'Vorgriechisch'/'Pelagisch': Neue Erwägungen zu einer älteren Substrathypothese. In Alfred Bammesberger & Theo Vennemann (eds.), *Languages in Prehistoric Europe*, 85–98. Heidelberg: Carl Winter.
- Tichy, Eva. 2004. *Indogermanistisches Grundwissen*, 2nd edn. Bremen: Hempen.
- Tischler, Johann. 1979. Der indogermanischen Anteil am Wortschatz des Hethitischen. In Erich Neu & Wolfgang Meid (eds.), *Hethitisch und Indogermanisch. Vergleichende Studien zur historischen Grammatik und zur dialektgeographischen Sprachgruppe Kleinasiens*, 257–267. Innsbruck.

Vennemann, Theo. 1984. Hochgermanisch und Niedergermanisch: Die Verzweigungstheorie der germanischen und deutschen Lautverschiebungen. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 106. 1–45.

Watkins, Calvert. 2006. Proto-Indo-European. In Anna G. Ramat & Paolo Ramat (eds.), *The Indo-European Languages*, 25–73. London: Routledge.

DRAFT