# The new Arboretum of Indo-European "Trees"

## Can new Algorithms reveal the Phylogeny and even Prehistory of IE[1]?

Hans J. Holm
Hannover, Germany

Abstract

Specialization in linguistics vs. biological informatics leads to widespread misunderstandings and false results caused by poor knowledge of the essential conditions of the respective methods and data applied. These are analyzed and the insights used to assess the recent glut of attempts to employ methods from biological informatics in establishing new phylogenies of Indo-European languages.

## INTRODUCTION[2]

In the last ten years, the easy availability of phylogeny reconstruction packages has led to a sheer arboretum of newly developed "trees" of Indo-European. Assessments range from total disapproval by most traditional historical linguists to enthusiastic trashy circulation by magazines and journals. We may at least note with pleasure that they demonstrate a strong public interest in the Indo-European Urheimat question.

The authors are proud to distinguish the main languages, what is no progress at all, since these results have been obtained by even the oldest methods (cf. Holm 2005 [3.1.1]). However, in the higher levels most 'trees' - often only 'binary topologies' - differ from each other, as well as from traditional views[3], or show only insignificant – brushlike - branchings. The reader is left with these differences unexplained, and parallel work is seldom analyzed. Thus, these new results are not beneficial.

Where are the reasons for these differences? All studies up to now preferred a 'trial and error' approach. However, it was too often difficult to distinguish whether the differences (or errors?) are due to the data or the methods, or both[4]. In this study therefore, we will analyze the data and methods applied, following other scientific reasoning:

- For the main "problem two" – subgrouping - we shall analyze the different methodological approaches and check whether the applied methods are appropriate for the subgrouping of languages. This final aim requires before,
- A look at the 'final' test options adduced in the two fields.
- According to this line of reasoning, we analyze the functional conditions and assumptions for which the adduced algorithms were designed, in particular, whether these are given in linguistics;
- As a basis we need to look at traditional methods of subgrouping in historical linguistics;
- First, let us start with the often involved easier (?) "problem one" – glottochronology:

---

[1] Indo-European

[2] I owe thanks for helpful comments and corrections from many sides, most of all Sheila Embleton, Joe Felsenstein, and Johann Wägele. Of course, all remaining mistakes are my own responsibility.

[3] E.g. presence vs. absence of the Indo-Iranian or Balto-Slavic group.

[4] cf. also Nakhleh et al. (2005).

# PROBLEM ONE – GLOTTOCHRONOLOGY

## Definition

Glottochronology is the computation of real-time in language history under the assumption of constant rates[5] of decay. It is a subfield of lexicostatistics[6] (Anttila, 1989, p. 396f). Glottochronology itself is a mechanistic approach, foreign to traditional historical linguistics and the humanities in general. Where, then, did the idea come from?

## Non-linguistic background

In the early fifties, the American linguist Morris Swadesh (1952, passim) heard of the $^{14}C$-dating for measuring the age of organic material. In a 'trial and error' approach, he designed an analogous method to estimate a "lexical half-life" and further the time at which any two languages should have diverged from a common proto-language. For this purpose he devised lists (soon referred to as "Swadesh-lists"), which had two distinct aims: first, the "basic" meanings (concepts) were chosen to have representatives in languages across different cultures (see e.g. Lohr, 2000, p. 211); secondly the vocabulary was assumed to be most resistant against borrowing and too much replacing (cf. Anttila, 1989, p. 231) and thus to be usable for testing deep time ranges. This second property – resistance to borrowing - has often been disproved (cf. e.g. Haarmann, 1990).

Only a few decades later, it was discovered that in nature these estimates, in addition to the stochastic scatter, underwent considerable variations[7]. The same is true in biology, where such changes arise in two forms: Horizontal (lateral) exchange / recombination, in the course of reproduction, which is not a change in the narrow sense, but rather a spread governed by selection. This exchange is normal in populations of higher species[8]. Real changes or "mutations*"* (replacements) are much less frequent. Many scientists with a mainly mathematical background assume a regular rate of changes here. However, already Fitch / Margoliash (1967, p. 283) found that, "Indeed, from any phylogenetic ancestor, today's descendants are equidistant with respect to time but not, as computations show, equidistant genetically." In view of these difficulties with reality, the methods are increasingly being differentiated, by grading down these overall rates to single species, genes, or even sites (characters). There seems to be a tacit pre-scientific belief in perpetual motion machines, avoiding the search for realistic environmental reasons, which would be inaccessible to mathematical rate computations. Such reasons have been found in the many internal as well as such environmental influences as e.g. radiation from the sun, which varies in time. There are even different areas of natural radiation on earth[9]. None of these is at all constant. Thus, an example for an extreme slow evolution is the 1938 rediscovered Coelacanth, existing nearly unchanged since 400 Million years (cf. Fricke, 1988).

Regrettably, all this has obviously been forgotten by the employers of phylogenetic reconstruction methods, as it excludes all methods requiring ultrametricity. Nevertheless, scientists keep on trying to compute time depths of biological evolution, and are now again transferring their algorithms into linguistics.

## Changes in linguistics are different

In contrast to biology, there are striking and decisive differences between languages and species: Language is a communicational system[10], and is thereby much more open to changes than any biological species, since languages are not genetically inherited, but learned. Textbooks usually list many types of language change and different views of the reasons for this. It is useful to put these into a chronological order of primary and secondary changes:

---

[5] "Rates" express a relationship of a variable to a constant time unit, e.g. velocity for changes of distance in time, e.g. km/h.
[6] This is in turn part of Quantitative Linguistics (cf. HSK-vol. 27, see Holm, 2005).
[7] Today the variations of 14-C-dating are "calibrated" by different measures.
[8] "A species is a population of interbreeding individuals that is reproductively isolated from other species." (Croft, 2000, p. 196)
[9] See e.g. L.Forster et al. (2002, p.13950-13954).
[10] Cf. e.g. Labov (1994, p. 9f).

*Primary changes from outside a language*

Primary changes are events and situations in human history, which is widely accepted by historical linguists: W. P. Lehmann (1980), for example, stressed, "… linguistic theory must regard language as an activity with a history." Raimo Anttila (1989, p. 391) followed, "In environmental factors the social ones are the most important, and social factors are of course further anchored in a particular society, …, and a particular historical event.", and Anttila (1992, p. 34) – after decades of linguistic scholarship – admonishes us that "History is theoretically primary in the matters of language and its use …" W. Croft (2000, p. 1) starts, "Language change is a historical phenomenon." Also, French linguists complain about the neglect of history, e.g. Jacquesson (2003, p. 117f)[11], «Les hommes construisent leurs langues plus qu'on ne croit souvent … . Ce phénomène explique l'échec de l'hypothèse < glottochronologique >.» and concludes, «… il existe des déterminismes, mais leur logique d'application est historique, elle dépend des événements.» And a German source (Schlerath 1992, p139): "In Wahrheit stellt die Herausbildung jedes einzelnen indogermanischen Sprachzweiges (wie z.B. Germanen, Kelten, Griechen, Inder) und die dann später erfolgende weitere Differenzierung und Ausbreitung der Sprachen, die zu dem jeweiligen Zweig gehören, ein völlig neues Problem dar. Diese Vorgänge können gänzlich verschiedene Voraussetzungen besessen und einen völlig verschiedenen Verlauf genommen haben."

What is the consequence? History never repeats itself; historical events are unforeseeable, in time as well as in intensity of their impact. It follows directly that history is not regular or constant. This we will call 'axiom one'. It implies that linguistic changes cannot occur at any constant rate in time and that therefore they must never be mathematically projected into the past or future. Simply by chance, the changes may vary around some peak(s) in a stochastic distribution or not even this (cf. Fig.1). Precisely this is the decisive mistake[12] in all 'glottochronological' attempts, which additionally compute time depths. Needless to say that changes might have happened in a very short time between long periods of very few changes (sometimes referred to as "punctuated equilibrium").

Anttila (1989, p. 179) sums it up in the hypothetical question, "If biological change can be largely characterized as coming from behind, by automatic natural selections, and cultural evolution from in front, from a conscious purpose, where does language fit in?"

The possibility to count the exchange of lexemes in whatever time for whatever written languages, and accidentally find (as M. Swadesh) languages having the same amount per time, cannot be a scientific proof, as soon as there is a single unexplained counterexample. Moreover, there are many of these, as has been demonstrated exhaustively: E.g. Labov (1994, p. 10), or Tischler (1973) made a fine case for that. For readers not familiar with the history of languages, let us additionally regard some more counterexamples of such impacts:

Of the IE family, language groups have lost very different amounts of the reconstructed original vocabulary, e.g. Germanic[13] as one of the better preserved ones has lost about 30 %, mainly by replacements from an unknown (pave Vennemann) substratum.

Albanian has lost about 80 % of its IE vocabulary[14] not 'by' time or an inherent urge, but - besides the influences of all the former and later neighbors - mainly through the superstratum of Roman domination. It should additionally be understood that these replacements as well as all other "distances" have nothing to do with the original genealogical relationship of Albanian or any other language, respectively. Educated scholars should in fact not regard such kinds of constellations as computable property of Albanian or of Albanian lexemes.

---

[11] I am obliged to Prof. St. Zimmer, Bonn, for this French source.

[12] The unpredictable decay of single radioactive atoms must not be compared with single socio-historical events of language change. That would mean to confuse the macro- with micro level. Moreover, it must be clear that 14C decay varies only within *certain limits,* (See above). This contrasts sharply with languages, which can change to any degree and in any time, as I am not the only one to have amply demonstrated.

[13] Cf. e.g. the counting of Bird (1982).

[14] Cf. Bird (1982); Haarmann (1990), passim; E.P. Hamp (2002, p. 682-3).

English replaced 50 % of its Germanic vocabulary not by time, but through Norman dominance after the Battle of Hastings, in addition to a long-lasting educational and clerical background of Latin. Albeit this had only a minor impact upon the 'basic vocabulary', this obviously did not happen in form of a 'rate'. The same holds for the six (in the 100-item list, or 12, in the 200-item list) of incorporated North-Germanic vocabulary, not in any rate, but in the time of the 'Danelag', through mighty Viking settlements. None of this constitutes any rate of change inherent as a property of English, as these authors must believe.

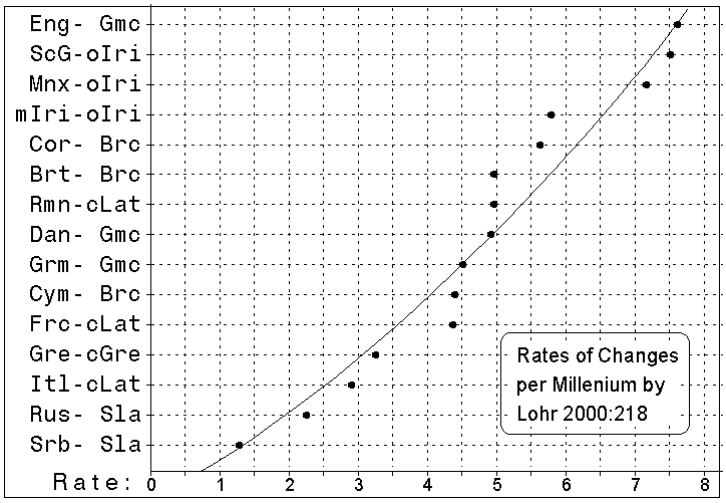



Fig. 1: Lexical Changes projected into "millennium rates"

Though 'basic' vocabulary is lesser affected as the rest of it, the differences are quantitative, not functional, and naturally, there is no principal and clear-cut border between these and other parts of the lexicon. Instead, we find a steadily changing distribution (cf. Zipf 1965). For M. Lohr (2000) has undertaken a tremendous effort to achieve an improved dataset, her computed millennium rates (Fig. 1), projected from changes between stages of some IE languages, may serve as a good example.

The hypothesis of glottochronology is thereby scientifically falsified.

In all these primary impacts described, linguists further distinguish different forms, the so-called strata (or layers). They work in three ways[15]: As adstrata (borrowings/loans), predominantly from prestige languages - which might sometimes even be own dialects[16] or older stages[17], substrata from subdued speakers, and superstrata from superior speakers. While borrowing is the main and often only concern of many researchers, the extremely variable impact of the latter two strata is often neglected.

In languages, communities of speakers themselves decide which of these primary changes they accept or not, under whatever physical or psychological 'pressure'.

*Secondary or language internal changes*

The primary changes, upon their full incorporation, most times end in disturbances of the original lexical, morphological, and phonological system. Accordingly, speakers reshuffle their system in ways, which seem to be universal to humans and have amply been described in textbooks of linguistics, social psychology, and psycholinguistics. These often proceed slowly and unconsciously, thereby obscuring the underlying historical events.

## Conclusion: No rates in language change

We could fill volumes with more examples[18]. It can be doubted whether there exists a culture-free basic vocabulary at all (cf. Campbell 1998, p.180f). Thus, we have to recognize that - at least in languages - there is not, and never has been, any inherent 'rate' to be projected into the past. To sum up: Though many addressed authors assume, require, or additionally work with some form of 'clock assumption', i.e. fixed rates of replacements per language or per meaning along computed edges in a hypothetical topology, their chronological

[15] E.g., French has a Gaulish substratum and then different (Gothic, Burgundian, and Frankish) superstrata; cf. e.g. Anttila (1989, p. 171); Polomé (1990, p. 331-8); Kontzi (1982), in general; Chap. V of Ernst et al., (2003).
[16] E.g. the London dialect at 'Early Modern English'.
[17] For details, see e.g. Lehmann (1992, p. 266ff).
[18] For another line of argument, see A.& R. McMahon (2000).

conclusions are therefore generally inadequate. For these reasons, we shall not further discuss any glottochronological attempt in detail.

## PROBLEM TWO: SUBGROUPING

### Definition and basic terminology

A phylogeny, in biological systematics, is a graph intended to represent genetic relationships between biological taxa. This comprises more than the woolly notice 'classification' without defined criteria. Linguists, in this respect, prefer to speak of 'subgrouping' of a language family.

Most methods in this review were originally designed for problems in biological systematics. Since the object of this study is however language subgrouping, we shall generally use the linguistic terminology, and make only a short comparison of the respective terms[19] in Tab.1, for readers coming from either field.

| Linguistic vs. biological data according to their origin, in … | | |
|---|---|---|
| Language Relationships | | Biological Systematics |
| 1. Homologous, inherited similarities | | |
| 1.1. Retentions (residues) of inherited original features | From remote ancestor | Symplesiomorphies, (not: generic traits) |
| 1.2. Shared (common) innovations | From last common ancestor | Synapomorphies, (shared derived traits) |
| 2. Analogous, not inherited similarities | | |
| 2.1. Borrowings, loans, copies, strata | Lateral shift | Horizontal transfer, diffusion |
| 2.2. Homonymies, chance agreements | Convergences | Homoplasies |
| 3. Differences | | |
| Individual replacements | Individual | Autapomorphies |

Tab. 1: Corresponding terminology

### Analysis of the problem

We start with the simplest case of two languages, as in Fig.2. Here arise two questions: first, are $L_1$ and $L_2$ at all related; and if so, in what directions? There is not a problem of subgrouping yet. This arises by adding a

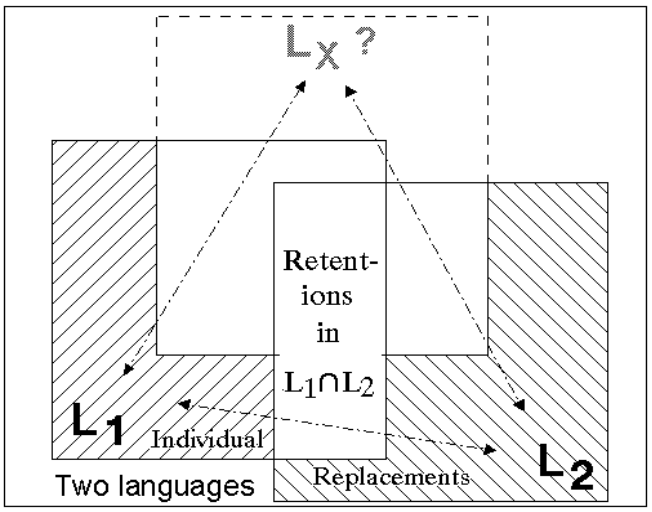
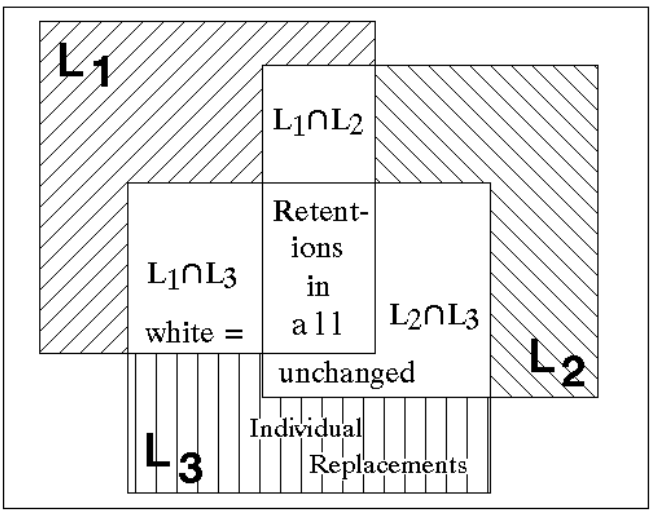



Fig. 2: Basic relationship: Blank interction = unchanged common features




[19] The terms in biological informatics used here go back to Hennig (1984, passim).

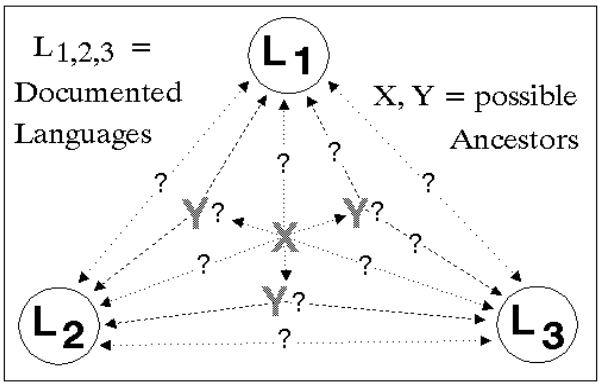



Fig.4: Three languages: Complete subgrouping problem

| Features | Languages | | |
|---|---|---|---|
| | L 1 | L 2 | L 3 |
| n1 | 1a | 1a | 1a |
| n2 | 2a | ? 2a | 2b |
| n3 | 3a | 3b | ? 3b |
| n4 | 4a | 4b | 4c |

Fig. 5: Reduced subgrouping problem

third language, as in fig. 3. We can easily distinguish individual replacements (line-fills), against the unchanged retentions (left white).

Here in fact the problems explode, if we regard all of them in Fig. 4.

Given that we exclude a mutual descent, and assume a former common ancestor, the problem can be reduced to that in Fig. 5.

We can clearly detect two kinds of features:

- Feature n1, shared by all languages, only proves that these languages are related. Feature n4, individual to each language ('autapomorphy'), only shows that these should be single languages. These features seem of no help to decide subgrouping and are therefore regarded as 'trivial'.
- Feature n2 combines $L_1$ with $L_2$ and would prove that they should be more closely related by a common ancestor 'Y', if there were not feature
- n3, pointing to another common ancestor, i.e. between $L_2$ and $L_3$.

These features are therefore regarded as 'non-trivial'.

The required decisions are qualitative, to be made exclusively by professional historical linguists. Only these decisions can and must be the basis for any quantitative approaches. Thus, historical linguists should regard quantitative approaches not as rival, but rather as complementary techniques. We shall therefore only recall those criteria of historical linguistics as needed to assess the validity of data used by the different authors.

## A short glance at the traditional criteria

*Linguists should have discarded trivial or irrelevant features.*

| Features | Languages | | | 'Outgroup' |
|---|---|---|---|---|
| | L 1 | L 2 | L 3 | L 0 |
| n2 | **Shared innovation?** (Synapomorphy) ./. X + | | **Retention?** (Symplesiomorphy) | ? |
| n3 | **Individual replacement?** (Autapomorphy) | **Retention?** (rest of symplesiomorphy) | | ? |
| | " | **Loans, Chance agreements** ? (Analogous similarities) | | ? |

Fig.6: Essential conditions

The essential shared innovations have to be kept apart from the following interfering ("phylogeny-uninformative") features:

One part of (the trivial) individual features would be individual replacements (autapomorphies), confined to one language and displaying neither an agreement with an outgroup feature nor structural analogy with the family. These in turn may indeed have destroyed former homologous features (retentions as well as shared innovations) down to remnants, then appearing as pseudo-autapomorphies.

Linguists will also identify the following trivial features:

- (Remnants of) retentions (symplesiomorphies), because they combine, rather than distinguish the languages and cannot reveal subgrouping. To identify these original features, biologists as well as linguists try to use so-called 'outgroups', taxa that are not part of the

family under study and thus prove common features between both to have preserved an older, common state. E.g., for decisions on Brythonic, Old Irish could be an outgroup (cf. e.g. Hamp, 1998, p. 313). Retentions may or may not show up in outgroups, because they could have been replaced or lost; but if they show up, this is a strong argument for a correct decision. Too often, we do not have this possibility: For IE, perhaps Afro-Asiatic (Hamito-Semitic) would do, but the available research[20] is not yet generally accepted. To choose ingroups as outgroups (e.g. Rexová et al. below) is inadequate. For other aspects of outgroups see later below. We shall not go into the detailed decisions of historical linguists because these should have been made before applying quantitative methods and is described in every textbook.

- Chance agreements (convergences, homoplasies), because they have nothing to do with relationship at all.
- Loans (Borrowings, copies, adstrata, cultural 'Wanderwörter', lateral gene transfer / shift) seem easily detectable, as they should contradict regular sound laws. Regrettably, this is not possible, if they have taken place before the particular sound shift occured. Thus, these 'historical loans' remain a problem, in particular if they occur only between neighboring languages, or are overlooked by unprofessional knowledge, as demonstrated below for the 'Dyen-list'. At least they normally are a sign of neighborhood in historical times.

*Linguists should have identified the 'shared innovations' (Fig. 5 and 6, n2).*

The most accepted criterion is the same in both sciences, called

- Shared or common innovations[21] in linguistics, and
- Synapomorphies in biological systematics, even
- Shared scribal errors in stemmatology[22].

As we have seen above, this identification is not at all easy. The difficulty of distinguishing the shared innovations sought after, from remnants of retentions, camouflaged by multiple replacements in all other branches, leads to the 'symplesiomorphy trap' in biology (cf. Wägele, 2001, p. 217-8), which is a danger in all methods. This is particularly difficult in the otherwise very desirable cases of pairs.

### Interim result

Yet, after following these 'perfect rules' over 200 years, linguists have not been able to reach agreement about the phylogeny of Indo-European. Many subgroupings are discussed again and again; in particular, the position of the Anatolian languages waits for a solution. Minor dissent can still be observed about the Italo-Keltic[23] or Slavonian relations and many others. Nevertheless, in defiance of these poor results[24], they are adduced as 'empirical proof' in some studies addressed below.

## APPROACHES FROM MOLECULAR SYSTEMATICS

### Why are these methods applied?

Some linguists came with the intention of updating the Swadesh approach; others wished to avoid leaving the field of quantitative evaluation to pure mathematicians. Some researchers, in particular biologists, appeared with the superficial view that linguistic change seemed to be similar to biological evolution, namely equaling the replacements of words with that in molecular material. In the first place, however, the easy availability of computer packages called for new fields. Let us look at the first aspect, the data:

---

[20] Cf. St. Georg (2004) on Nostratic attempts. In order to limit space, I shall not cite the special sources on Hamito-Semitic etymology.
[21] Cf. e.g. Porzig (1954, p. 55); Hamp (1992 & 1998, p. 307ff), with additional criteria; Croft (2000, p. 15); Ringe et al. (2002, p. 66).
[22] Descendence of manuscripts.
[23] Where possible, I write the unambiguous 'k' for the respective phoneme, here along with the authority of Hamp (1998) and others.
[24] Cf. Hamp (1998, p. 342).

## Do properties of the employed data meet the assumptions required by different methods?

*Do linguistic properties correspond to biological data?*

In mathematical analyses since the second half of the former century, biologists have increasingly worked with DNA-sequences of operational taxonomic units[25] (OTUs) like the following

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| OTU 1 | A | A | T | C | G | T | A | C | A | G | G |
| OTU 2 | A | A | G | C | G | T | . | G | A | G | G |
| D | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |

Tab. 2: With 3 changes in 11 sites, yielding a 'Hamming' distance[26] = 0.2727.

Besides this, protein-frequencies are in use, as well as morphological or ethological material[27]. Molecular sequences of supposed related taxa must be 'aligned', that means, elements brought into the corresponding positions ('site', 'locus') in the gene or chromosome to be compared. A special difficulty for this alignment procedure is 'gaps' (insertions and deletions, e.g. the missing 'A' in OTU 2 above). Sequences differ in their degree of variability; well preserved ones (e.g. the Ribosomal RNA 16S) are suited for deep studies, not for recent, fine grained phylogenies; at the other extreme are the highly variable sequences, which allow differentiation even of individuals, and are therefore used in forensic or family studies. This insight should also be taken into account for quantitative language comparisons.

However, such dissimilarities, in contrast to biology, if regular, define genealogical relationship rather than distance. A resembling procedure in linguistics is cognation, i.e. analyzing whether forms could be homologous or not, e.g.:

```
German : haupt|und, compared to
English: hea.d|and, as an entire word - meaning relation.
```

The few studies on phonological 'Hamming distances' have found only little interest in linguistics (detailed reasons follow later). Moreover, phonological data often behave differently in respect to lexical ones, which we shall mention in the following paragraphs. They must not be mixed up with other data in the same method, because the methods must be tailored according to the properties of the employed data. Phonological data are only used by Lohr and Ringe et al. (see below), the latter additionally with morphological material.

The biological function of such sequences is often not yet known. Thus, lists with functions (~'meanings') as variables[28] are seldom used in molecular systematics. Moreover, biologists are aware that morphological functions may change extremely under environmental pressure, concealing genetic relationship (e.g. the fins of whales). Now biologists reckon to have found similar data in linguistics in form of word-lists. Again *contrasting to biology*, the meaning / communicational function is well known in languages. This leads to the dichotomy of two types of data used in lexicostatistics:

| Meaning | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| German | a | b | d | e | g | i | k | l | n | O |
| English | a | c | d | f | h | j | k | m | n | P |
| Distance | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |

Tab. 3: Hamming-distance between meaning lists, here D=0, 6

(1) In lists from the onomasiological[29] point of view, meanings are taken as 'characters' (coded as numbers in Tab. 3) and their (multi)nominal representations in the languages[30] under study as the character states or forms, coded e.g. as small letters. Thus, most of the researchers referred to work with comparisons like this, apparently resembling molecular sequences, coded e.g. as small let-

[25] Term for variables under study, e.g. (groups of) languages or dialects, species, genome sequences 'S', the 'leafs' or tips in a 'tree'.
[26] Named 'observed' distance 'D' in PHYLIP (Felsenstein, 2004a), elsewhere also 'apparent', 'p(henetic)'distance, whereas Swofford et al. use 'p(ath)' for its length in a tested topology.
[27] Cf. e.g. Wiesemüller et al. (2002, p. 59ff).
[28] Comparable with linguistic "meaning lists".
[29] Problem of how concepts/meanings are named. Cf. e.g. Anttila (1989 [7.2]).
[30] Onomasiological lists or dictionaries, as e.g. Buck (1949), continued for European languages by the late Schröpfer (1979, passim).

ters. Thus, most addressed researchers work with comparisons like this, apparently resembling molecular sequences.

These lists are often referred to as Swadesh-lists[31]. One of the many representatives of these is I. Dyen's Indo-European list (1997), which three teams (details later on) made use of. For other language families the temptation is great to obtain such data from any dictionary or other source. This can easily lead to errors, e.g. since it is well known in linguistics that meanings are much more prone to variations than forms are, often completely concealing the original meaning or even swapping to the contrary (cf. Middle Engl. sely 'happy' → Mod. Engl. silly).

(2) In lists from the etymological point of view the probable reconstructions[32] of the original forms serve as characters and the presence or absence of homological derivates or 'cognates' in any language as the binary states or values (cf. e.g. Bird, 1982). Most historical linguists regard these lists as a much better choice for comparison and reconstruction, for the following reason, "Die Grundlage muß immer die materielle Identität … bleiben. Sie behält ihre Tragfähigkeit, auch wenn die Funktionen größere Divergenzen aufweisen. Das Umgekehrte ist nicht haltbar und kann nur zu unbegründeten Annahmen und Verwirrungen führen[33]." (O. Szemerényi, 1990, p. 30). These lists can best be obtained from professional etymological dictionaries, which provide the highest reliability. Up to now, this type has only been used for the SLR method (addressed later).

In both types of lists, the character states have been chosen to be 'cognate'**.** Cognacy (homology) is established by relatively reliable sound laws. 'Reliable' here means that they should be supported by some complex cognates (in biological terms a sufficient 'homology frame') and of course, contradictions are explained. Note that etymons represent the inherited retentions, in biological terms pure 'symplesiomorphies'.

If these are not assessed by historical linguists, being the best experts in the etymology of the languages under test, there is a high danger of errors in the classification (coding) of cognates, in particular between retentions (symplesiomorphies) vs. shared innovations (synapomorphies), loans (adstrata), and chance agreements. A typical case is the above-mentioned list of I. Dyen. Native speakers, as brought in by one team, could easily fall victim to so-called 'folk etymology'.

The next feature occurring in both systems is synonyms (polymorphism). In biology, 'polymorphism' means the parallel existence of two or more states at the same site. These states could consist of all types listed in tab. 1, often e.g., plesiomorphies mixed with apomorphies - not always distinguishable from the former - and arise in the form of alleles, which, in the course of evolution, would either be abandoned or fixed. If both are homologous, they may be divided into two variables. In languages, polymorphism shows up in the form of different expressions ('synonyms') for one notice ('meaning') in dialects, levels of speech, or other niches, e.g. labor vs. work in English. These cases should be solvable by a narrower definition. This could be done by weighting; e.g. the etymon ie. *kuon 'dog' appears as <Hund> in German with the original meaning, but as <hound> in English in a specialized meaning only, the original being covered by 'dog'. In etyma lists, deviating meanings can be much more tolerated. Of course, the 'devil is in the details' here, because the original meaning of an etymological construct was not necessarily the simple intersection of recent semantic features.

Back changes (reversions). In biology, backmutations (e.g. A→T→A) between the only four nucleotides per site are much more frequent than between the 20 amino acids; less frequent are back changes of morphological characters. These problems are addressed by many biological methods. However, they hardly arise between languages, where undetected back changes of whole lexemes are extremely rare (cf. e.g. Seebold, 1981, § 234;

[31] In fact, there were different ones by Swadesh alone, cf. Embleton (1995, p. 267), with references. Additionally, there exist at least a dozen attempts at improvements.

[32] Usually marked by a preceding star or asterisk.

[33] 'The basis must always remain the … material identity. It keeps workability even when the functions show greater divergences. The reverse cannot hold and only leads to unfounded assumptions and confusions.'

Ringe/ Warnow/ Taylor, 2002, p. 70), and can be neglected. Here again we encounter one of the often overlooked differences between linguistic vs. biological change.

Chance agreements (homoplasy). In biology, homoplasy occurs[34], caused (1) by chance parallelisms, or (2) convergence, based on environmental pressure and leading to morphological similarities which in fact are not homologies (e.g. carnassials' teeth in marsupial vs. recent tigers), or "homologies in the wrong phylum[35]" (Sudhaus & Rehfeld, 1992, p. 111). Between languages, we have to distinguish the levels: Chance agreements (homonymies) of words /lexemes occur rarely between basic vocabulary lists, at 2 - 8 % depending on their complexity (see Holm, 2005 [3.1.2] for a survey of the relevant sources). Of course, they are most likely to arise between short words.

In contrast to molecular sequences, phonemic systems as the building blocks of words are much more complex than e.g. the only four nucleotides. Phonemic developments differ in frequency depending on the level of observation, too: Diachronically, we can easily observe phonemic variation, more often between vowels than consonants. In the speaker community, these changes remain often unconscious, and do not necessarily change the status of homology, or establish another language (cf. the 'great English vowel shift'). Similar phonological changes/ sound shifts occur in different languages of a family, e.g. the merger of PIE a with o > a in Grm, Bal-Sla, Ind-Ira, and Hittite. The Kentum: Satem border, regarded as so significant in former times, has very much lost its importance (cf. Tischler 1990). Of course, they also may happen among languages far apart and not closely related, e.g. p > f in Germanic, as well as e.g. in Iranian, or Arabic. Cf. further the '$k^w$'-splits in Keltic as well as in Italic.

Phonological data must therefore be regarded as trivial universals and less suited for genealogical studies.

Strata (gene shift). Last, horizontal (lateral) gene transfer (shift) in biology normally only happens intraspecific[36], exceptions appearing only between lower organisms not dependent on reproductive restrictions. In languages, lateral shift, also called: 'adstrata', simply 'borrowing', or 'interference' (cf. Porzig, 1954, p. 53f, Croft 2000, p. 145ff) is frequent, but can often be ruled out by sound laws. Even higher amounts of changes may arise from sub- and superstrata.

*Stochastic properties*

The other features, which are seemingly 'trivial' or 'non-informative', are individual replacements (autapomorphies), isolated to a single branch (OTU). These, in the first place, are of course indicators of a single language or species.

However, they are in no way 'trivial', for they have a stochastic side effect: They can and do contaminate not only former retentions and shared innovations to individually different, often major extents, thereby determining the amount of dissimilarities, the distance, between any pair of languages. E.g. in Fig. 7, nine of the originally 15 common replacements in language $L_2$ have been destroyed by the subsequent 25 more individual chance replacements. In the worst case all of our common innovations — traditionally needed for the recognition of sister languages — could be lost. This ends up in the 'proportionality trap' (Holm, 2003). For a short explanation, let us look at Fig. 7:

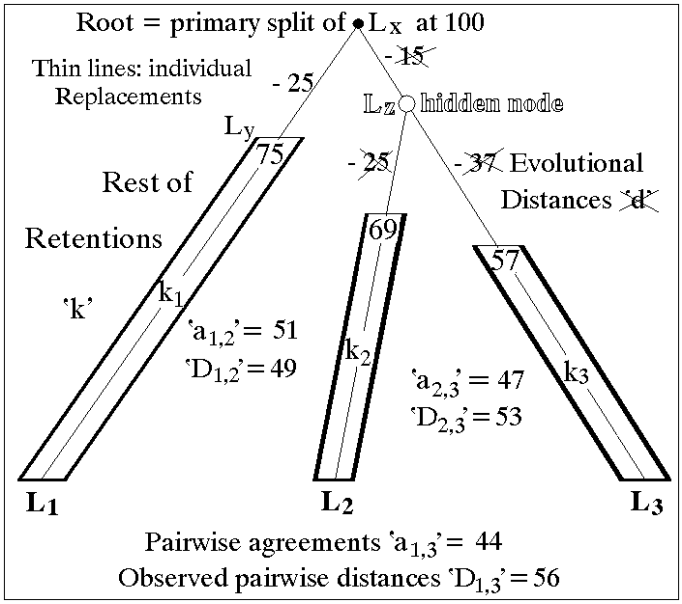



Fig. 7: Distances and agreements in three languages

[34] E.g., up to 25% in the case of the four nucleotides, one of which will be replaced by any change.
[35] In technical terms, 'a phenetic similarity contradicting a phylogeny.'
[36] Within the same species.

Let language $L_X$ with assumed 1oo original features split into its daughter languages $L_Y$ and $L_Z$. Subsequently $L_Y$, by 25 individual replacements, becomes the recent language $L_1$, where thus 75 original retentions $k_1$ are left over. The second daughter, $L_Z$ has first lost 15 original elements by replacements or innovations. It then splits into recent languages $L_2$ and $L_3$, starting with these 15 shared innovations. However, they subsequently undergo further, individual replacements, affecting not only the original retentions, but also – by chance - parts of the 15 innovations, and this in different amounts[37]. Having understood this, the results should no longer puzzle us: We can clearly observe that the second split (at 100 original features minus15 innovations) can no longer be defined by the number of agreements 'a' in the daughter languages $L_2$ and $L_3$, since $L_2$ exhibits more agreements with $L_1$ instead with its closer relative $L_3$. In particular, we cannot detect this hidden node at 85 in the distance or intersection of agreeing cognates between languages two and three.

In Fig. 7, we can clearly recognize that the amounts of agreements 'a' between any pair display only a superficial resemblance or similarity. This is only by chance and not a logical, representation of the proportional genealogical relationship between these languages, as erroneously claimed by the methods carried over from biology.

This resembles the so-called 'problem of long edges[38]' in biological phylogeny reconstructions, arising from multiple replacements, which produces two consequences: first, the substitution of essential former innovations, and secondly, mainly in molecular systematics, with the increase of replacements, increase analogies and reversions.

Thus, working with plain (dis)similarities, can lead to at least incompatible, wrong phylogenies[39]. This means that at least observed distances[40] between linguistic features are not 'additive' in principle. Views like "... languages that share more recent common ancestors tend to be more similar than languages with more distant ancestry." (Pagel, 2000, p. 189), exhibit a typical case of this proportionality trap, since the amount of similarities between any two languages can only be a measure of their genealogical relatedness, if both languages

- started/parted with exactly the same amount of original features N; and
- developed with exactly the same rate of decay, ending in the same amount of retentions 'k'.

### *Phylogenetic properties of distances*

Phylogenetic additivity requires the distance between any two taxa / languages to equal the length of the path in the phylogeny between them. In practice, this is never the case, at least between languages, as can clearly be deduced from Fig. 7.

Ultrametricity is even more restrictive, for it requires a so-called evolutionary clock', i.e. glottochronology, which is neither given in biology nor in language, as we have already seen. Therefore, methods employed under this assumption yield unacceptable results (see Nakhleh et al., 2005 [5.2]).

### *Direction, polarity and 'rooting'*

Though traditional as well as quantitative methods assume[41] an original, oldest ancestor, which has to be found, it can sometimes be doubted whether a tree with one root only exists at all. E.g., Croft (2000, p. 196) admits, for "mixed" languages, "… they do have multiple parents, contrasting to the family tree model." At least, the tree model is often an over-simplification[42] in the subgrouping of languages. If M. Pagel (2000, p. 189) claims that "... languages**,** like biological species, evolve in a predominantly hierarchical manner..." this is of

---

[37] Of course, the intersection of agreeing cognates 'a' between $L_2 \cap L_3$ cannot be 57, because these retentions are situated at different places ('sites' in genome sequences) of $L_2$ and $L_3$.

[38] Alternatively "long branch attraction", cf. Swofford (1996, p. 427); Felsenstein (2004, p. 120f).

[39] Again, even if strongly supported by high bootstrap values, which only test the consistency by amounts of supporting features.

[40] In biological systematics, this effect is known as the difference between observed phenetic 'D-distances'. Quantitative phylogeny tries to transform these into evolutionary 'd-distances'. This is not possible in language research.

[41] or define, e.g. 'Indo-European'

[42] Cf. e.g. Aikhenvald (2001, p. 4ff).

course correct, but might conceal the decisive difference, namely - unlike higher biological species - the persistent ability of exchanging any amount of features after (!) a split.

In this respect it is unsatisfactory that all these programs from biosystematics, which are not based on the false 'ultrametric tree assumption', e.g. Maximum Parsimony (MP), yield 'unrooted' trees, or better, topologies. This might reflect reality in dialect or population analyses. Nevertheless, if we assume and wish to detect, an evolutionary process, we need some idea to find the direction[43]. The researcher therefore has to look for other methods to find the starting point. Generally, the following methods are employed:

'*Outgroup*[44] *comparison'*. Most researchers use this technique, albeit in different senses. In simple cladistics, this is managed mechanically by the addition of a single outgroup. Without analysing the characters to distinguish between apo- and plesiomorphies, major errors are unavoidable here. Reductions in outgroups lead to wrongly assessed apomorphies (Sudhaus & Rehfeld, 1992, p. 111). Better results can be obtained by a-priori analysing the characters, including comparisons of as many outgroups as possible (cf. Wägele, 2001, p. 178). In linguistics, this principle is also known and applicable, as already set out.

Some biologists as well as linguists regard *complexity* of a feature or synapomorphy in biological evolution as sign for a later state in evolution, because it would take more time to develop. However, many exceptions are known in both fields: In biology, e.g. snakes do not have legs, but never represent any primitive state of the tetrapods they belong to. In languages, we can also observe both opposing trends. Often speakers tend to replace complicated words, grammar, or phonemes, by simpler ones (cf. e.g. the loss of morphological features in modern English). This can be observed every day in the case of people acquiring foreign languages, e.g. migrant workers, as well as in first language acquisition by children. Thus, this argument is of little help in finding a root.

*Paleontology* (the evaluation of extinct taxa) is valuable in both biology and linguistics; but these species or languages have certainly also undergone replacements / autapomorphies after their split-off, so that not every feature can automatically be regarded as homologous. The unprofessional employment of such data, as e.g. by Forster and Todt (2003) is not accepted by historical linguists [45], as well as the use of only recent data, as (albeit deliberately) in I. Dyen's list.

*Ontogenesis* can be used in biology to trace archaic evolutions[46]; not so in linguistics, where language acquisition by children is pure learning, and e.g. the development from one- to two- to multi-word speech does not help in any way in rooting language family trees.

### *Interim result: No advantages so far*

It must be clarified that all these lists only allow distinguishing equal or different entries between any two languages. Most authors of the lists employed have only tried to identify 'cognates' by meeting the sound correspondences (in one case not even this). The following methods are designed to solve difficulties we do not or seldom have in languages, as e.g. back changes (reversions), large amounts of chance agreements (homoplasy), or rooting problems. Moreover, they cannot use the decisive criteria of traditional historical linguistics, namely shared innovations. Let us see what they offer instead.

## 'Distance' methods

### *General*

'Distance methods' use the sums of dissimilarities between any two taxa (cf. 'D' in Fig. 7). To my knowledge, these methods from biological systematics were first recommended to linguists by M. Ruvolo (1985, p. 193ff), but not applied to natural languages.

---

[43] Cf. e.g. Felsenstein (2004b, p. 6).

[44] Taxa / languages that do clearly not belong to the 'ingroup' under study.

[45] E.g. Szemerényi (1990, p. 7ff); Meier-Brügger (2000, E509); or Seebold (1981, §40, 322); in particular Eska/Ringe (2004).

[46] Origin and development of individuals, which are assumed to reiterate their phylogenesis. For seldom counterexamples, cf. Wägele (2001, p. 180).

The old distance methods (hierarchical agglomerative, or cluster analysis) used up to the 1980s have mostly become obsolete, because they distorted the data and did not guarantee an optimal tree. Even a full plot of the original data by hand reveals the raw similarities better and is not at all difficult, if one starts combining every language separately with its next and next-but-one neighbor (see Fig. 8).

These old methods did not take care of different amounts of replacements, where the observable amounts of distances then are naturally smaller than the actual (evolutional) ones. "The extreme of this view is the phenetic perspective in which it is asserted that nothing but the extent of similarity matters biologically ... ." (Swofford et al., 1996, p. 487). Well-known are the four distance-input methods offered in the PHYLIP package (Felsenstein, 2004a), applied by two teams addressed below, despite the underlying assumptions that are not met in language change.

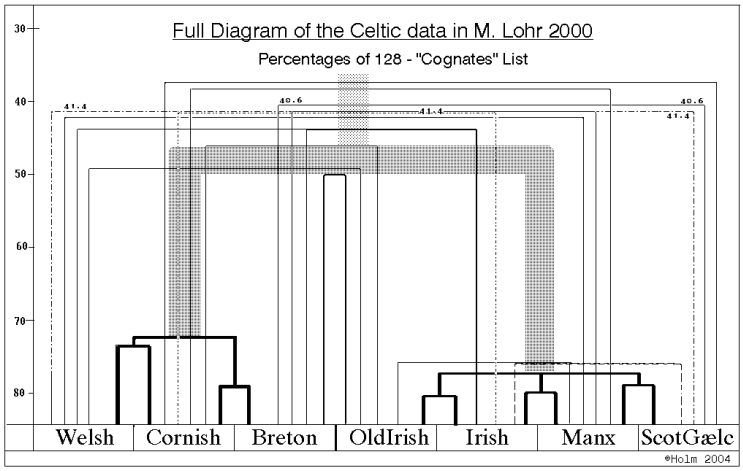



Fig. 8: Phenetic full plot by percentages of observed cognates with no distortion

This applies in particular to the two programs thereof requiring the already described *ultrametric condition:*

- The first program is the 'Unweighted Pair Group Method with Arithmetic mean' (UPGMA) of the 'NEIGHBOR' sub-package. UPGMA naturally distorted the Keltic data in Lohr (2000, p. 213), and yielded false results, while a full plot as e.g. Fig. 8 displays the data correctly, but probably - due to the proportionality trap - not the correct tree. Needless to say, the Ringe team (Nakhleh et al., 2005 [5.2]), repeatedly pushing against this open door, made the observation that "… UPGMA did clearly the worst with respect to both criteria[47]."

- The second one is the heuristic search program 'KITSCH' of the PHYLIP package.

The other two programs yield unrooted trees and require the somewhat looser condition of *additivity*. If that is not the case, both programs require prior transformation, for, "… (they will not make a statistically inconsistent estimate) provided that additivity holds, which it will if the distance is computed from the original data by a method which corrects for reversals and parallelisms in evolution." (Felsenstein, manual Distance Matrix Programs, version 3.6).

- The heuristic one here is the 'FITCH', which could reconstruct all possible trees, where the data can be corrected by the sub-options 'Fitch and Margoliash' (1967), 'least squares', or 'minimum evolution'.
- The 'Neighbor-Joining[48] (NJ)' program of the 'NEIGHBOR' sub-package transforms the distances into quasi-ultrametric ones (cf. Swofford et al., 1996, p. 487ff) by using the arithmetic mean to all other ones. It is sensitive to loss of shared innovations and to the sequence of taxa fed in. It "... is guaranteed to recover the true tree if the distance matrix happens to be an exact reflection of a tree." (Felsenstein, 2004b, p. 166). At least between languages, this seems not to be the case (cf. Holm, 2003), and the method should be abandoned, what is also concluded by the Ringe team in Nakhleh et al. (2005, p. 21).

Moreover, replacements in different languages do not simply vary a little around some mean, but in any degree, as amply exemplified in the chapter on glottochronology above.

[47] Compatibility and "established aspects of IE history".

[48] The term seems somewhat misleading, for here a search algorithm by star-decomposition is meant, obviously unaware of synonyms in the older neighbor-joining procedures of hierarchical agglomerative clustering methods.

ASSESSMENT: All distance methods use only one parameter between languages, namely the observed distance 'D', ignoring the stochastic dependence of 'D' on the three further parameters (see Fig. 7). This essential relationship is only employed in the SLR method (addressed later). It follows that all the results are necessarily prone to this basic error. Even after transformation of the distances, only a statistically consistent 'tree' is expected, what must not yet be the correct one[49]. Moreover, it seems to be generally accepted that distance methods are outperformed even in biology by likelihood methods (cf. e.g. Swofford et al., 1996, p. 446).

## 'Character state' methods

The following methods evaluate sequences (or meaning lists respectively) character by character. They assume more or less regular replacements of words, or mutations of alleles. Since these character state methods often depend on individual decisions, individual errors in the data must necessarily lead to wrong or self-contradicting topologies. If highly complex data are treated in the same way as simple ones in one algorithm, non-significant ones or convergences can incidentally outnumber the former, highly informative data. We must bear in mind that meaning lists do not distinguish between retentions vs. shared innovations, since both are homologies: the former from the earliest, the latter from the latest common ancestor.

The first method, under the "*Maximum Parsimony" (MP)* criterion assumes the topology with the shortest evolutional path (i.e. the lowest sum of replacements) to be the best guess. Regrettably, this old hypothesis in biological evolution alone bears only unreliable evidence of the genealogy of languages, because the latter has nothing to do with the minimum of replacements (cf. Holm, 2003): We must recall that English remains a West Germanic language in spite of the high amount of Romance replacements. Additionally, MP naturally tends to yield inconsistent results when faced with very different, in particular long peripheral branches (cf. e.g. Felsenstein, 2004b, p. 117; Mount, 2004, p. 248, 251; Swofford et al., 1996, p. 427,494), which is precisely the case with some IE data[50]. Moreover, MP yields exhaustive results only up to at most 20 taxa. For the 85 taxa of e.g. the Dyen-list, 'hill-climbing heuristics' must be employed, which in fact cannot guarantee best results. Since MP is therefore principally unsuited for this list, and no sub methods are cited in Rexová et al. (2003), who alone employed it, we will not go into more detail here. Unaware of these shortcomings, Nakhleh et al. (2005) retested the method on their own dataset.

The *Maximum Likelihood (ML)* heuristics are nowadays suggested as a better choice[51], because they account for different replacement chances. As for all probability algorithms, a stochastic process is presupposed, where replacements have been constant, independent, homogenous, undirected, and reversible. Since this is never the case even in biology, the shortcomings are met by explicit 'model' settings for different distribution, evolutional behavior in preferences and rates of replacements of the four nucleid acids[52]. Because such models are not feasible for languages, no team has used any of the available programs (in spite of claiming to do so). Instead, a likelihood-related method, *Bayesian Inferring (BI),* is used. It allows for site heterogeneity, but still erroneously assumes replacement rates, by inferring 'posterior' probabilities from 'prior' distributions in hypothetical topologies (cf. Felsenstein, 2004b, p. 288ff).

The so-called "perfect" *Warnow Compatibility Method (CM)* used with Ringe (1995, passim) chooses the highest compatibility of subsets with multinominal data. This appears to be a perfect optimality criterion. This method is recommended in biology, when the rate of evolution varies among sites (e.g. Mount, 2001, p. 248). It strictly requires characters to be uniquely derived. However, this requirement is too often destroyed in languages as already described above. Additionally, by relating on the pure amount of compatible characters, the already mentioned impact of extremely different amounts of replacements after a split has not been taken into account. If

[49] Note however that the results may resemble reality if *by chance* the environmental circumstances are not too far from the conditions of the methods and/or the signals are strong enough.
[50] In particular Hittite, Albanian, and English, which naturally then behave 'recalcitrant'.
[51] E.g. Swofford et al. (1996, p. 528).
[52] So-called 1-, 2-, 3-, or 6-clock assumptions (cf. Swofford et al. (1996, p. 434); Wägele (2001, p.232,267).

additionally the data fed in are not assigned as innovations vs. retentions, some strange results are not in the least astonishing.

The *network approach*[53] was originally derived from the split decomposition method (Bandelt & Dress, 1992, passim). It visualizes the surface structure of the data as they are, including contradictory features as reticulations, but without a root. This could be inferred from the full diagram (cf. Fig. 8), what on the other hand soon becomes too confusing with more than a dozen branches. Of course, both must be called phenetic, because they cannot distinguish between superficial and even false similarities between features. Two more teams employed the method in newer work. A 'network' program, applied to the simple two-split problem of Fig. 7, ended in a star graph, detecting neither the primary nor the secondary root.

### Interim Result

We have to thank the Ringe team (Nakhleh et al. 2005) for a trial series of these methods (network and SLR excluded) using their improved dataset. However, the reader is not told that Anatolian is used as a standard outgroup to achieve rooting of the topologies, and normally not an outcome of the methods. Moreover, they leave us with the different outcomes unexplained. Remarkably, no method accounts for the most acknowledged principle of common innovations or synapomorphies, albeit they could easily be employed as allowed only once to appear on the tree. Regrettably, the team refused a test of the following approach:

## APPROACH FROM STOCHASTICS

### Idea and rationale

In 1950, during a discussion following some proposals on IE subgrouping at the Research Section of the Royal Statistical Society[54], the well-known British statistician D. G. Kendall remarked: "We must not expect to be able to determine this [the epoch of separation] as a date in history, but we may hope to be able to construct a statistic, large values of which will imply an early, and small values a late, epoch of separation." What is the rationale of that 'statistic'?

### Separation-Level Recovery (SLR)

Let us assume an ancestral or 'mother'-language $L_x$ with N characters, which will - in no fixed rate any - decrease in time. Let $L_x$ split into two daughter languages $L_i$ and $L_j$ at node $L_y$ with a common amount of $N_y$ features. Then both will – after due time and independently of each other - replace different amounts of features, leaving different rests '$k_1$' and '$k_2$' of inherited ones. Naturally, there should be left some agreeing inherited features 'a'. Now: Note that these do not vary around some rate, and additionally are stochastically determined by the three parameters N, $k_i$ and $k_j$, what is not at all conceivable by ad hoc 'common sense' and has therefore been easily overlooked. We assume now that a linguist has analyzed these languages, found them to be genealogically related, and has determined the cognates in them. He can then count the number of these cognates '$k_i$' and '$k_j$' and the number 'a' of agreeing ones.

However, he does not yet know, when these languages parted, in particular, if they have parted earlier or later than other related languages. Now - a knowledge of the nodes $L_{y=1,2,3,\dots}$, determined by the estimation of their amount of features $N_y$ at the era of split (as already described by Kendall above), would give us a rank of departures.

Because this problem, to detect the unknown node $N_y$, is just the reversal of the historical events, it can be solved by the *hypergeometric*[55] estimator:

---

[53] This is sometimes classified as a distance method. Nevertheless, it works character by character, and distances are the output.

[54] Of November 25th, 1949, published (Kendall 1950, p. 49), but never since cited. This is the reason why this author was unaware of this approach when detecting these relations through working on Indo-European material of Bird (1982).

[55] For detailed proof and explanation cf. Holm (2003)

$$E(N_{y=i,j}) = k_i\, k_j / a_{i,j}$$

E.g., in Fig. 7, we can only in this way estimate the node '$L_Z$' at

$$E(N_{z=2,3}) = 57*69 / 47 = 83{,}4 \sim 85.$$

These nodes can then be visualized by different heuristics described in Holm (2005). This method also assumes a stochastic process of replacements, but neither constant nor reversible.

This extremely important stochastic feature is an inherent property of language change, and to a smaller extent of biological evolution, too. It contrasts sharply with the narrowed assumption of minimal evolution or rates of replacement in biology[56].

**Interim result**

The method is able to estimate the original amount of homologies independently of any later contamination, even if there is no single innovation left. The algorithm is robust, as chance agreements only influence the result as divided by the amount of common agreements. A disadvantage is the stochastic scatter of the estimations, which can only be distinguished from bad data by the sophisticated logical methods analyzed in Holm (2007a). 'Bad data', technically termed 'systematic bias', arise e.g. when they are too heterogeneous by their semantic fields. This would end in different chances of replacement between the respective lists and consequently false stochastic results. It follows that instead of a large dictionary of unknown heterogeneity, a small, but perfect list (of about 200 reconstructions), would be the better prerequisite. Perfect here means quantitatively and qualitatively complete decisions of cognacy in all the languages under study. The advantage of this method on the other hand is the ability to detect the root and stages of separation. The method has been amply used in biology for capture-recapture research, but never before for inferring phylogenies of species.

## TESTS OPTIONS

Tests in mathematical systematics are limited to *data robustness*, which falls precisely into the proportionality trap, since larger amounts of supporting data would automatically resist better to the random errors implied. Some teams are proud of presenting high so-called 'bootstrap-values'. These arise if we have many shared features for a branch, which then cannot easily be destroyed by random test changes. The test naturally gives poor results if there are only few shared features. Moreover, the test cannot detect chance agreements (homoplasies) and borrowings. Almost the same holds for the so-called 'jack-knife test'. The 'Bremer-Index' reveals the amount of agreements for a subtree; thus, it is subject to the same weaknesses.

*Empirical* tests of plausibility in biology comprise comparisons with competing methods, historic-biographical patterns, and other classes of data.

In *linguistics*, the most commonly cited criterion is the agreement with established aspects of IE history. Ringe (cf. Nakhleh et al., 2005) choose "Indo-Iranian, Balto-Slavic, and the further eight main branches." This is in fact only a minimum requirement met by nearly all methods. Following Hamp (1998), I would like to add Italo-Keltic. Further, the resulting phylogeny should completely and without contradiction be transformable into *real geography***,** starting from a 'staging area' (Urheimat), reconstructing the paths of migrations or expansion into the recent or oldest known seats. In fact, there are dozens, if not hundreds of views on Indo-European origin and subgrouping (cf. e.g. Day, 2001 or encyclopedia articles). So far, there have been only fragmentary visualizations. Languages should then be checked for borrowings in the neighbors encountered along these routes.

[56] Changes in biology would most times not fulfill the conditions for this hypergeometric distribution, as $k_i + k_j$ should exceed 0,2N, where acceptable spread can only be expected with above 0,9N.

# APPLICATIONS TO INDO-EUROPEAN DATA

## Based on a homespun phenetic list (35 meanings)

*One recent attempt is the 'Celtic list' by Forster & Toth (2003)*[57].

AIMS: F&T intend to throw light on the relationship between Keltic and IE, in respect of subgrouping as well as glottochronology.

DATA: They reckon to use the paleontological aspect, by evaluating some sparse Gaulish material, as well as a dozen native speakers. Choice and cognation are totally based on *superficial resemblances,* without any help from a professional historical linguist[58]. Why did they not at least resort to the etymological dictionaries available for different Keltic languages? Additionally, the input data seem to be insufficient because of the scatter, which can only be made up for by lists significantly exceeding 100 variables (cf. Lohr, 2000, p. 211; McMahon / McMahon, 2002, p. 25; Holm, 2005). Moreover, these few data were additionally reduced (thereby increasing the uncertainty) by allowing only an upper bound of five states per variable.

ROOTING: Claiming to compensate for this negligently accepted shortcoming, they performed a so-called "negative proof" by a Basque list, supposed to be genetically unrelated. The authors simply took five 'spurious identities' as a measure of possible chance agreements. Aside from the fact that a few linguists assume a Basque substratum in Indo-European, there are additionally many loans into Basque (e.g. from Latin and even Keltic[59]). Simply to carry over these undefined overall error figures to all other pairs is a forbidden generalization and inadmissible between languages (cf. Holm, 2005 for a survey of other studies on chance agreements).

METHOD: The "Network" approach.

ASSESSMENT: It is an inherent feature of this method that even single characters acquire the status of decisive criteria for a split or reticulation. However, being able to visualize contradicting traits does not involve finding the true ones. Moreover, because "The linguistic network approach is therefore expressly intended to search for treelike structure in potentially 'messy' data.", this intention fails, since significant homologies are not detected and decisions are made by linguistically insignificant variants.

The results, e.g. separating Gaulish from 'Insular Keltic' are not accepted by many linguists. Neither is the brush-like split between Latin, Greek, and Keltic. A detailed discussion has meanwhile been written by Eska / Ringe (2004[60], p. 569-82). This - like all mathematical methods - simply mirrors the input mixture of wrongly with correctly assigned states[61]. The additional glottochronological attempt once more implies – as demonstrated at the start – senseless computing of history.

## Based on Dyen's "Swadesh-type list" (207 meanings)

The subsequent three teams used this list, obviously because of two apparent advantages: The data set is readily obtainable from the Internet, and seems easily convertible for the involved programs because of its electronic coding. Available are the raw data and two distance matrices. The latter (IE-PERC84 or 95) contain, between every language, the decimal fraction of n1/(n1+n0), where n is the number of determinable cognations, n1 positive (i.e. 'cognate'), and n0 negative. The sum n1 + n0 = all determinables, seldom reaches the 200 of the list, because of assumed questionable data or decisions. It follows that these numbers must not simply be treated

[57] The publication in PNAS is astonishing, since this is neither read by linguists nor evaluated by linguistic search engines.
[58] Forster naturally holds that this is detectable by reticulations; but what, then, is correct?
[59] E.g. izoki(n) 'Salmon', cf. Pijnenburg (1983, p. 240).
[60] Eska & Ringe professionally criticized the data of F&T, and, less convincingly, the glottochronology, but the network method with only poor understanding. The following clash (Language 81-1/2005, p. 2-3) made this even clearer.
[61] "Garbage in - garbage out" (old programmer's wisdom).

as 'observed distances[62]', but decisions about handling the 'gaps' and questionable data are required, a question which no researcher here even addressed[63], let alone the distributional properties.

The data, prepared as long ago as 1960, are much less reliable as compared with those of Ringe et al. (2002). In the current Internet version (of 1997), for 'English ST' alone, at least seven loans were still erroneously coded[64] as cognates, while the 12 borrowings from French are correctly assigned (cf. Embleton, 1986, p. 100; 1995, p. 266; Thomason & Kaufman, 1988, p. 265ff). No team has noticed the review by Tischler & Ganter (1997), only one of them the one by Embleton (1995, p.266). The mistakes admonished by her have not been corrected up to now. They have already led to a false position of English in Dyen et al. (1992) and again in two studies addressed below (McMahon / McMahon and Gray / Atkinson). There seem to be more mistakes, if one only starts with the first meaning 001 'ALL', where Dyen lists Alb. 'GJITHE' as autapomorphy[65], ignoring Pokorny (1959) and Demiraj (1998)[66]. Loans are continuously coded as "cognate only in this class" (e.g. of dialects), e.g. 046 FEW, Alb. 'PAK', which is of course a loan from vlat. paucu.

In contrast to current views (Cowgill, 1986, p. 64; Anttila 1989, p. 305; Hamp, 1998), all three methods dependent on these data connect Romance with Germanic, instead of with Keltic.

Two teams additionally transposed the original data into a secondary binary matrix. Linguistically, by this procedure the original list with a few flaws might get even more biased: Only one of many, many more examples is e.g. the Keltic representation for Dyen's item "066 HAND", which is regarded as original IE retention (cf. Pokorny, 1959, p. 805) and has cognates in nearly all other IE languages with just slightly different meanings, as e.g. 'palm (of hand)', but would appear as only one positive mark (under Germanic) in this type of binary list.

The first team having used this list was

*Rexová/Frynta/Zrzavý (2003) – RFZ,*

who display a special understanding of linguistics: Of course, Swadesh (1952) is not "the earliest quantitative lexicostatistical method ..." at all (see e.g. Embleton, 1986). With this background, the authors believe they are entitled to solve even the IE Urheimat problem en passant.

DATA: Dyen's list. They claim to use "individual cognate classes suggested by the linguistic methods ... (sensu de Pinna, 1985)". De Pinna (reference missing) is no authority in linguistics to be invoked here. In fact, they only used Dyen's "cognition classes", additionally reduced by 141 character states, to meet limitations of the applied PAUP-package (Swofford, 2002). Astonishingly they must have missed the matrix Dyen provides, since they state that they have converted the list themselves.

METHOD: RFZ claim that "... no explicit optimality criterion has been used by the comparative linguists." without the least knowledge of even the basics: In fact, it has always been clear in Historical Linguistics that the relationship between languages is e.g. closer by the criterion of more shared innovations (in particular morphological ones), and not a shortest evolutional path in a topology, as they tacitly accept by using the MP approach. It must still be up to the specialist in his or her own branch of science to declare the criteria, and not the mathematician dabbling in unknown terrain. Turning to 'the traditional lexicostatistical' approach (obviously restricted to the views of Dyen et al., 1992, and confusing 'lexicostatistics' with glottochronology), they disqualify this as "phenetic[67], based purely on general similarities. Thus, cladistic methods, which do not use professional knowledge of the nature of their material, must be called "phenetic" themselves (cf. Wägele, 2001, p. 178). To

[62] After being converted from the original agreement percentages, of course.
[63] Dyen, Kruskal, & Black (1992) did address this question.
[64] The relatively complicated coding of different types of "cognation" obviously deters specialists from reviewing the decisions.
[65] in his words, "cognate only in this class" here Albanian internal.
[66] Via www.indo-european.nl
[67] 'Phenetic' means 'judging after (superficial) phenomena', i.e. things that appear but the cause of which is in question.

come to the core: The authors subjected the Dyen list in three transformations to "the MP[68] approach", which is not the 'state of the art' in molecular biology, and moreover, is unsuited for languages, as shown above.

ROOTING: They employ Hittite as an outgroup, simply by copying the weakly based assumption of Ringe et al. (2002).

ASSESSMENT: Out of the 85 taxa employed, we are informed about the 11 main branches only. The three outcomes are self-contradictory, e.g. the position of Albanian naturally changes from one 'tree' to the other, hardly agreeing with conventional views. This is of course due to the very low number of residues in Albanian, which are then extremely sensitive to any error. One former mistake is the "altered multistate list" with intentionally inserted negligencies, another the derived binary matrix (p.122) with 2,456 states. That view is only strengthened by the unsolved different outcomes – in particular for Albanian - from different coding here. These empirical observations, in addition to the methodological assessment, should be sufficient to show that the approach is inadequate. The stochastic conditions for the percentages are not recognized, a matter, which can in no way be healed by the assumption of shortest evolutional paths.

*Gray & Atkinson (2003, passim),*
to my knowledge, are the last in this series.

AIMS: This work is explicitly aimed at the obsolete *divergence times, namely of Anatolian.*

DATA: The authors also exploit the Dyen-list, unaware of its intentions and shortcomings (Atkinson & Gray 2006, p. 93: "contains expert cognacy judgments"), extended it by Hittite and Tocharian words of then unspecified origin[69] and also (as RFZ above) transform the list into a secondary binary matrix of here 2,449 character states. On p.436, we are told that the initial coding procedure makes no allowance for missing cognate information, what might have caused some bias, since gaps in fact appear in every variable of that list, e.g. 78 lists contain 12.5 gaps (Dyen et al., 1992, [2]).

METHOD: G&A seem to be the most 'modern' from the geneticists' point of view, as they employ the free MrBayes-package (Huelsenbeck et al., 2001) that takes care of the different frequencies of the character states. The principle is mathematically as clear, as the connections to linguistic change are unclear. The package is supposed to apply the Bayes theorem, i.e. to compute the posterior probability (of a tree) from a prior probability under the likelihood of given data. A very critical point is the model chosen: A time-reversible model with parameters 'frequency' and 'substitution rates', where in particular the latter cannot be accepted. Under this model, the Marcov-chain-Monte-Carlo (MCMC) method generates changes on the tree. The tree is then chosen that fits "most likely" to the binary data. By the way, all other methods are claimed to fit the data best. Thus, it would be interesting to know why this method should do better. Instead, the authors in every article prefer to sledge the dead horse of the Swadesh's algorithm.

ROOTING: This method, too, yields unrooted topologies. Again - as in the Ringe attempt - cursory or non-specialized readers are misled about the position of Hittite. On p.437 (top), they cite two sources as "considerable support for Hittite ... as the most appropriate root for Indo-European ...", e.g. Rexová et al. (2003) (sic!), obviously unaware, that these explicitly merely copied Ringe et al. (2002). Of course, no linguist would regard Hittite as a "root of IE". The other source is even lesser accepted.

GLOTTOCHRONOLOGY: This aim, declared[70] as the main point, should in fact make us suspicious. The time estimations were deduced from the claimed knowledge of 14 'nodes', then naïvely projected back with the aid of the 'Markov Chain Monte Carlo' method, 'Penalized Likelihood Optimization procedure', 'General Time Reversible substitution model (GTR)', and 'Gamma distribution[71]', where the impressive (albeit correct) jargon terms do not at all guarantee them to be adequate. E.g., the assumption of the GTR that mutations are reversible

[68] Obviously in the default setting of PAUP v. 4.0b4a (Swofford, 2000).
[69] Now published in Atkinson & Gray (2006, p. 93).
[70] G&A deny doing glottochronology, because they are narrowed to the Swadesh method.
[71] All integral components of the MrBayes package (Huelsenbeck & Ronquist, 2001, passim).

does not apply to languages. They mean to reject the well-founded warnings of the Ringe-team against computations of divergence dates by quoting Ringe (2002, p. 61), where he in fact 'scored an own goal'.

ASSESSMENT: Let us start with only some examples of *language subgrouping*: English, due to the errors in the list, appears as an early offspring from W-Germanic, instead being a member of it, opposed to (Old-High)German[72]. There remain the odd results in the higher levels, e.g. the position of Italic with Germanic, clearly contradicting the result of Ringe et al. (2002, p. 112), where one wonders, how G&A can pretend (p436r) that "Recent parsimony and *compatibility* analyses also supported these groupings." This, where they themselves had just correctly cited that "Maximum likelihood methods generally outperform … parsimony". In test applications to the data in Nakhley et al. (2005), "MrBayes", contrasting to all other tested methods, always combined Keltic with Germanic. Further, the Slavic group appears quite different from mainstream opinion represented in e.g. Campbell (1998), e.g. regarding the western group. Also completely opposing all traditional results, is the grouping of Albanian with Indo-Iranian, and RFZ got this result only in one of their three self-contradicting versions, namely in the secondary binary conversion. Here, the same data manipulation in both might have resulted in the same error. In Indic, the combination of Marāthī with Gujarātī contradicts all results of historical linguistics, due to late convergence and a typical outcome of Dyen's choice of data. The same holds for Dyen's "Afghan", what does not exist. He obviously speaks of Dari, a SW-Iranian language, belonging - contrary to their result - closer to Persian than to Waziri, which is a Pashto-dialect of SE-Iranian.

To support the senseless *glottochronological* outcomes, the authors (p.435) call upon the "Kurgan Theory" letting the Indo-European invaders start "beginning in the sixth millennium BP", mistakenly[73] referring to Gimbutas and Mallory. They next tell us that "... the Anatolian theory claims that Indo-European languages expanded ... from Anatolia around 8,000 - 9,500 years BP.", recurring on Renfrew (2000), who in fact (p.415, 419) gives 7000 BC. As already mentioned, there are several dozen more hypotheses and models of an IE expansion. A number of serious linguists, above all the late Larry Trask, have sharply attacked this attempt. In a "2nd Response to Trask", Gray & Atkinson[74] accuse him, "Trask displays a serious misunderstanding of biology." Perhaps, but they miss the point, which is subgrouping of languages. Here, as well as in pre-history, the authors display a serious misunderstanding of the character of language change and prehistory. G&A "... argue that there are a number of similarities that enable us to use phylogenetic techniques from biology to resolve questions in historical linguistics." Precisely these claimed 'similarities' - in fact superficial juxtapositions - do possess crucial different functional properties (as set forth above), not recognized by the authors. We must not mechanically apply methods from one discipline to another without profound understanding of the inherent functional and causal relationships. The linguist cited in proof (L.Campbell) cannot be expected to have a better knowledge of molecular biology and mathematics than the attacked L.Trask, and thus is simply not in a position to recognize these differences. Towards the end of the same document, the authors enlighten us, "...- we are not arguing about when the wheel was invented (we know [sic!] it must have been around 6,000 BC), ... ." According to latest calibrations, there is definitely no evidence for wheeled transport before c. 3637-3337 cal BC[75] (cf. Fansa/Burmeister, 2004). In Atkinson & Gray (2004, Table 16.2, 3) this claim is merely repeated, not substantiated. On p.293f they again try to teach the late L.Trask now the history of the IE word for wheel: Here they maintain that the wheel-word should have been borrowed some 6,000 years ago (contrary to their first figure) long after the era of 9,800 - 7,800 BP[76], which they computed for the IE divergence. Moreover, G&A not only in

[72] Note that there are different terms for the subgroups, where sometimes OHG belongs to another "W-Germanic" group.

[73] Gimbutas (1992, p.6) gives 4400 – 4300 BC for her first wave, corresponding to 6400 – 6300 calendar years, or seventh millennium ago.

[74] http://www.psych.auckland.ac.nz/psych/research/Evolution /Response%20to%20Trask%20 Take2.doc.

[75] The difference represents the uncertainty of the 14C determination plus the 'wiggle' areas of the calibration curve.

[76] Unaware of the technical definition of "BP" in archaeological science, they seem to mean "sun-years ago", thereby referring to 7800 to 5800 BC, when in fact agriculture expanded from Asia Minor.

this article demonstrate serious ignorance about the etymology of the words for wheel and wheeled transport in IE. Thus, they have no linguistic proof at all.

It remains a riddle, from where they took the prior estimates for the Hittite lexemes, in particular, for this is the core statement of the whole article. Above all, location of the 'Urheimat' goes far beyond the evidence of their method. This work regrettably reached a wide audience by being published in 'Nature[77]', which is a journal of natural sciences, with no competence in humanities.

*McMahon/McMahon (2002, passim)*

DATA: After earlier attempts with phonological data, MM detected that Dyen list, and described it p. 29 as a "... distance matrix, which is based on the percentages of non-cognate forms between each pair of languages." This is simply not the case, as described above. They close with a very detailed excursus on meaning lists, where they state on p.50 that, "... this greater resistance to borrowing has never really been tested, ..." Obviously they are not aware of e.g. Haarmann (1990), who addresses this question, particularly testing Latin loans in Albanian basic vocabulary, or e.g. Viberg (1983) and Wilkins (1996) who focus on regularities in linguistic change in 'basic vocabulary'. A. & R. McMahon (2002, p. 47), also ascribe errors to "the Swadesh 200-word list" (in fact speaking of the Dyen list), but inconsequently make not the least effort to correct these errors. Admittedly, this is an extremely tough task, additionally complicated by the coding of the data.

METHODS: They announce p.29, "that there are computer programs which draw and select the most parsimonious tree." In fact, there are, but the ones they used do not belong to this category (cf. e.g. Felsenstein, 2004b, p. 133). Instead, they employ three distance heuristics available in PHYLIP, neglecting that the preconditions are not met. Their first outcome is "... from the Neighbor program." That "... the PHYLIP programs ... are selecting from the population of possible trees ..." (p.29) is again wrong regarding that first option. The second method employed is 'FITCH' under the additive tree assumption - which is not given between languages. They go on p.30 (also p.47), "... the Maximum-Likelihood approach of the Fitch and Kitch [sic] programs ...", which is not true, either (see above, and the manuals of PHYLIP[78]). They also used 'KITSCH', under ultrametric assumption, but of course, the outcome is only reported in passing. As announced in 2000, they later (McMahon/McMahon 2005) switched to the network approach. The results, in particular as represented on p.102 (here rooted by using Albanian as the outgroup!) equal those in the original Dyen (1992, Fig.1), as already suspected in Embleton (1995, p. 265): (1) The nonexistence of an Indo-Iranian group, by only a 1%-distance to the 'root', (2) Slavonian, without its tripartite division, with the position of Slovenian as nearly an outgroup to the rest, (3) the position of English as an outgroup (!) to the rest of Germanic, and (4) the position of Gujarātī with Marāthī to Indo-Aryan, due to unrecognized common borrowings, e.g. from Sanskrit. In addition, Romance is grouped, albeit insignificantly, with Germanic, instead with Keltic. Further, Provençal is grouped between Walloon and French (!).

ROOTING: The authors defend these unrooted trees as reflecting "... the acceptance in biology that all species ultimately derive from a common single ancestor". It consequently follows that there must be a root: Evolution happens in time and therefore is chronologically directed, and the common ancestor farthest back in time should be the root, as Proto-IE in an IE phylogeny.

ASSESSMENT: The authors claim that their methods "...identify the subgroups which would typically be proposed for Indo-European". In fact, even the worst methods of the last 100 years (see Holm, 2005) most times detected one or the other primary group, as far as the input was correct. The competing attempt of Ringe et al. (later below) is just briefly mentioned.

In general, all three attempts miss the axiom of shared innovations, and simply classify the languages by misuse of phenetic surface (dis)similarities.

---

[77] This journal again typically has no reviewer in the field of historical linguistics. Only April McMahon, in a later article (Nature, Science update, Nov. 18th, 2003), regrettably remarked, "This kind of study is exactly what linguistics needs."

[78] Of course, PHYLIP offers ML-programs, e.g. DNAML.

## Based on improved meaning lists

*M. Lohr (2000)*

AIMS: She aims at a rehabilitation of lexicostatistics and even glottochronology. When Lohr (2000, p. 209) writes about discrediting voices on glottochronology, "However, it is hoped that this chapter will suggest ways in which the perceived shortcomings of the method may be reduced, ..." she misses the point: Not the methods, but the underlying rate assumption is erroneous.

DATA: Suspecting that bad results from former methods could additionally be due to insufficient "basic vocabulary lists", she devises[79] a new 128-meaning list for 18 European languages of – regrettably - only five IE branches, diligently tested against five super phyla for even more universal stability. However 'stability', as she herself remarks (p.210), yields just smaller and thereby less significant amounts of replacements (cf. e.g. Kessler 2005, p. 65f), thereby competing against the benefits. The mutual cognacy percentages are presented in a matrix (and partly visualized by this author in Fig. 8).

OPERATIONALISATION: Lohr claims to have conveyed her similarity percentages into distance measures by taking their negative logarithm. Perhaps intended to take care of multiple replacements, this algorithm at the same time exponentially overweights lower distances, in particular under 0.4.

METHODS: These data are then fed into heuristic programs designed for distance data in the PHYLIP 3.5 package. First, she tries UPGMA, where she naturally - due to the not met ultrametric requirement - observes errors in the outcome. Then, she tries the 'FITCH' heuristic under the 'least squares' option, which also distorts the data[80]. Last tested is Grimes' and Agard's method on phonological data, where she observed that, "… the phonostatistical method … is inferior in several ways to the lexicostatistical one." Naturally (see above). Finally, she tries to compute history ('rates'), where her own results, as displayed in Fig. 1, speak clearly against glottochronology.

ASSESSMENT: Because of the limited candidates, the results cannot completely be compared with the minimum requirements in the above chapter 'Test Options'. Correctly, though nearly insignificantly, the Keltic languages show a common root with the Romance, and the Slavic with the Germanic. Thus, these results are better than those of MM above, which is more likely due to the correction of the data than to her tremendous work on the choice of the data set.

*Ringe/Warnow/Taylor (1995, passim)*

DATA: RWT place much emphasis on establishing their own professional[81] reliable (meaning, or character) list, in terms of historical linguistics: The data contained ca. 333 lexical characters of the 24 oldest known IE languages, as well as 15 morphological and 22 phonological features. The employment of *phonological* data is astonishing, since I am not the only one to doubt their relevance (cf. Lohr above and even Ringe et al. (2002, p.66) himself!).

OPERATIONALISATION: In contrast to most others, RWT work with characters, rather than with differences. However, the types of homology are alluded, but mixed up, e.g. when RWT (2002, p. 71) write, " ... each state of the character ought to represent an identifiable unique historical stage of development - a true homology.", symplesiomorphies are confused with synapomorphies.

METHOD: The *optimality criterion of maximum compatibility (MC)* rests completely on the narrowing assumption that languages are closer related the more features are preserved after a split (cf. the enumeration RWT 2002, p. 86ff). This is only another appearance of the proportionality trap. Its application with the presumed

[79] We are not told how the cognations were found.

[80] The formula presented by Lohr on p. 214, as taken from the manual to the 'FITCH' Program in PHYLIP 3.5 (improved in V 3.6 of July 2004), is not the one of the least squares option.

[81] However, contrary to scientific rules and e.g. the Dyen list, these data have never been completely published.

"perfect phylogeny" method seems to be dated meanwhile, because it is too sensitive to multiple replacements, and no longer used even in biology (cf. Kim & Warnow, 2004).

In Nakhleh et al. (2005), the team presented some interesting comparative trials under their dataset with the above addressed methods, except the network and SLR approaches. However, the reader is left with many unexplained differences, where the reasons are again sought in "incompatible data" (see below). The newer methods, as announced in the 2007-URL[82] cannot be assessed up to now.

ROOTING: The decision to choose Hittite as the 'outgroup' rests on admittedly questionable assessments for only two morphological traits ("M3, M5") as original IE symplesiomorphies, which could be rather central innovations as well. Precisely this questionable decision has been copied by two other groups (cf. above), but without noticing the caveats by Ringe (2002, p. 97f) himself.

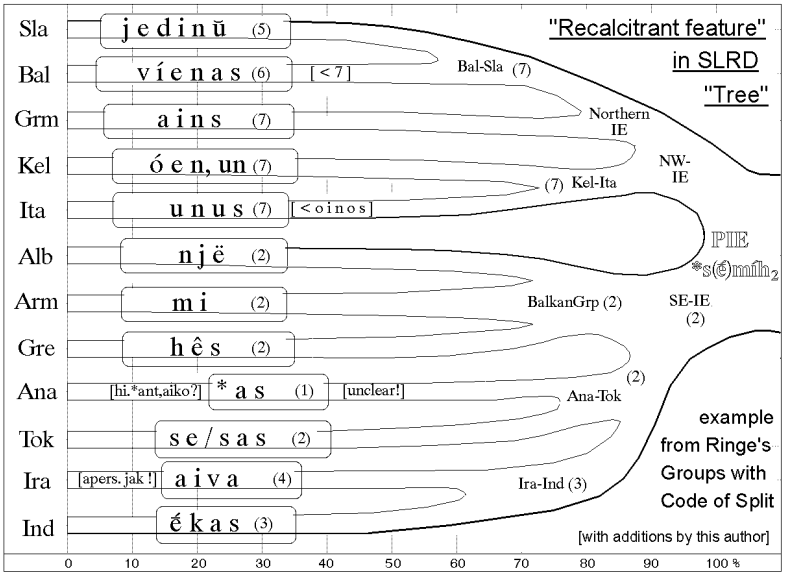



Fig. 9: Test of an "incompatible character" on the simplified SLRD IE phylogeny

ASSESSMENT: The trees presented are in fact far from being "perfect": Even on the 2007-web page, *Albanian is still grouped with Germanic, and Balto-Slavonian with Indo-Iranian*, both far from mainstream opinion. Already the first one of the "recalcitrant" characters, IE *smíh$_2$ (RWT 2002, p. 75), fits perfectly into the IE tree of the SLRD (based on Holm, 2007b). Reasons may be that the compatibility method decides after the maximum of agreements, that common innovations are not regarded, in spite of acknowledging this as primary, and employing phonological data in spite of criticizing their value himself. The former false position of Old English between (the satem groups) Indo-Aryan and Balto-Slavonian in an often-cited web page[83] is obviously given up now. Nevertheless, the team tries to rule out undetected borrowing by involving network methodology. I am happy to observe that in a current web page of Nakhleh a new tree comes much closer to Holm (2007b), except for the position of Anatolian and Tokharian.

## Based on Rix et al., LIV-2 with 1195 etyma; Holm (2000, passim)

AIM: Infer the subgrouping of main IE branches, using the true *stochastic interdependencies* between all four determining factors of the cognate data.

DATA: Since the algorithm depends on a dictionary ordered according to etymological reconstructions, and containing only retentions (symplesiomorphies), the only available one for Indo-European then was the Pokorny (1959)[84]. As soon as the "Lexikon der indogermanischen Verben" (LIV-2, Rix et al., 2nd ed. 2002) appeared, this was used, for several reasons: Verbs are much more resistant against replacements than nouns; the team with all modern resources at a department of Indo-European should be more reliable, and last not least, much better data were available for Anatolian and Tocharian languages.

OPERATIONALISATION: The Pokorny was employed in form of the binominal list provided by Bird (1982), the LIV-2 was coded by this author in the same way.

[82] www.cs.rice.edu/~nakhleh/CPHL/#software
[83] "Our latest results suggest that it falls somewhere within the Satem core!" The URL changed in 2004 and was cancelled in 2005.
[84] The University of Leiden project of an update is still far from being completed.

METHOD: The 'Separation Level Recovery' (SLR), as outlined above. After a bias had been detected, assumed to be the reason for the odd late splits of poorly documented languages, the method was extended to account for the different distributions in the data, thus now named *SLRD* (=Distribution) as already used in Fig. 9 (cf. Holm, to appear 2007b). Note that the percentages refer to the residues left at the era of split, what could have happened in different groups at different times. Further research suggests that the IE expansion can be modeled better in a circle-explosion model, by which the duration of multiple connections can be displayed on real-map / time conditions, rather than by an only one-dimensional tree (cf. the slide show via www.hjholm.de).

Empirical ASSESSMENT: The outcome perfectly fits Fig. 15-2 in Anttila (1989, p. 305), including the grouping of Albanian with Armenian. All major groupings of the different Ringe teams are of course recognizable, including the Italo-Keltic relationship, agreeing also with Cowgill (1986, p. 64), and Hamp (1998, p. 342).

## RESUME AND OUTLOOK

Scholars searching for parallels between biology and linguistics have to take into account the *many* differences in the fields and levels. It has been amply demonstrated that languages behave significantly differently from biological species. In biology, nature varies according to more or less constant environmental influences. These variations are then selected by survival conditions. In the humanities, the causality is vice versa: The human brain created language as a means of communication and can change it (or not), according to the needs and fashions of the social scenario in history - in no rates in time ever. Therefore, future 'phylogeny' research must not simply apply methods designed for biological data, which cannot exploit the knowledge base of historical linguistic specialists, in particular, the distinction between the different origins of the features. Moreover, not only borrowings have to be included, but also the effects of sub- and superstrata, which contaminate the usability of distances or agreements by the criterion of the shortest evolutional path.

## REFERENCES

Aikhenvald, A.Y. & R.M.W. Dixon. (2001). *Areal diffusion and genetic inheritance*. Oxford: University Press.

Anttila, Raimo. (1989). *Historical and comparative linguistics* [CILT 6]. New York: John Benjamin's.

Anttila, Raimo. (1992). Historical explanation and historical linguistics. In Davis, G.W. & G.K. Iverson (Eds), *Explanation in Historical Linguistica*, [CILT 84], Amsterdam/Philadelphia: John Benjamin's, 17-39.

Atkinson, Q.D., & R.D. Grey. (2006a). Are accurate dates an intractable problem for historical linguistics? In Lipo, C.P., O'Brien, M.J., Collard , M., & St.S. Shennan (Eds), *Mapping our Ancestry: Phylogenetic Approaches in Anthropology and Prehistory.* New Brunswick: Aldine, 268-296.

Atkinson, Q.D., & R.D. Grey (2006b). How old is the Indo-European Language Family? Illumination or More Moths to the Flame? In Forster, P., & Renfrew, C. (Eds), *Phylogenetic Methods and the Prehistory of Languages.* Cambridge UK: McDonald Institute, 91-109.

Bandelt, Hans-Juergen, & A.W.M. Dress. (1992). Split decomposition: The new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* I-3, 242-252.

Bird, Norman. (1982). *The distribution of Indo-European root morphemes*. Wiesbaden: Harrassowitz.

Buck, C.D.. (1949). *A dictionary of selected synonyms in the principal Indo-European languages*. Chicago: University of Chicago Press.

Campbell, Lyle. (1998). *Historical Linguistics*. Edinburgh: Edinburgh University Press.

Cavalli-Sforza, L.L. & Edwards, A.W.F. (1967). Phylogenetic analysis models and estimation procedures. *The American Journal of Human Genetics* 19-3,I, 239-240.

Cowgill, Warren. (1986). *Indogermanische Grammatik.* Halbbd. 1. Einleitung. Translated by A. Bammesberger and M. Peters. Edited by M. Mayrhofer. Heidelberg: Winter.

Croft, William. (2000). *Explaining language change: an evolutionary approach*. Harlow: Longman.

Day, John V. (2001). *Indo-European origins: The anthropological evidence.* Washington DC: The Institute for the Study of Man.

Demiraj, Bardhyl. (1997). *Albanische Etymologien. Untersuchungen zum albanischen Erbwortschatz*. Amsterdam, Atlanta: Rodopi.

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

Dyen, I. (1997). Comparative Indoeuropean database collected by Isidore Dyen. *www.ntu.edu.au/education/langs/ielex/IE-DATA1*.

Embleton, Sheila M. (1986). *Statistics in historical linguistics* [Quantitative Linguistics 30], Bochum: Brockmeyer.

Embleton, S. (1995). Review of 'An Indo-European classification: A lexicostatistical experiment' by I. Dyen; J.B. Kruskal & P.Black. TAPS Monograph 82-5, Philadelphia. (1992). *Diachronica* 12-2, 263-68.

Ernst, G., M.-D. Gleßgen, Ch. Schmitt, & W. Schweickard (Eds). (2003). *Histoire linguistique de la Romania.* [Handbook of linguistics and communication science, vol. 23.1]. Berlin, New York: W. de Gruyter.

Eska, Joseph F., & Don Ringe. (2004). Recent work in computational linguistic phylogeny. *Language* 80-3, 569-582.

Fansa, Mamoun & St. Burmeister. (2004). *Rad und Wagen: der Ursprung einer Innovation; Wagen im Vorderen Orient und Europa.* Mainz a. Rhein: v. Zabern.

Felsenstein, Joseph. (2004a). *PHYLIP [package of programs for inferring phylogenies]*: Seattle: Department of Genetics, University of Washington.

Felsenstein, Joseph. (2004b). *Inferring phylogenies.* Sunderland MA: Sinauer Ass.

Fitch, Walter M., & Emanuel Margoliash. (1967). Construction of phylogenetic trees. *Science* 155, 279-84.

Forster, Peter & A. Toth, Alfred. (2003). Toward a phylogenetic chronology of ancient Gaulish, Keltic, and Indo-European. *Proceedings of the National Academy of Sciences of the USA* 100, 9079-9084.

Forster, L., P. Forster, S. Lutz-Bonengel, H. Willkomm, & B. Brinkmann. (2002). Natural radioactivity and human mitochondrial DNA mutations. *Proceedings National Academy Science*, USA 99:2121, 13950-13954

Fricke, Hans. (1988). Coelacanths: The fish that time forgot. *National Geographic* 173-6, 824-828.

Georg, Stefan. (2004). Review of S. Starostin, Etymological dictionary of the Altaic languages. Leiden: Brill, 2003/4. *Diachronica* XXI-2, 447ff.

Gray, R.D. & Q.D. Atkinson. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 406 [6965], 435-439.

Hamp, Eric P. (1992). On subgrouping and Armenian. *Annual of Armenian Linguistics* 13, 55-62.

Hamp, Eric P. (1998). Whose were the Tocharians? Linguistic subgrouping and Diagnostic Idiosyncrasy. In Mair, V.H. *(Ed), The Bronze Age and Early Iron Age Peoples of Eastern Central Asia,* Vol. 1, 307-46. [JIES Monograph 26], Washington DC: Institute for the Study of Man.

Hamp, Eric P. (2002). Albanian. *The New British Encyclopedia*, 22, 682-3.

Haarmann, H. (1990). 'Basic' vocabulary and language contacts; the disillusion of glottochronology. *Indogermanische Forschungen* 95, 1-37.

Hennig, Willi. (1984). *Aufgaben und Probleme stammesgeschichtlicher Forschung*. Berlin, Hamburg: Parey. (Engl. version available).

Hock, Hans Henrich. (1991). *Principles of historical linguistic*s. Berlin, New York: Mouton de Gruyter, 2nd ed.

Holm, Hans J. (2000). Genealogy of the Main Indo-European Branches Applying the Separation Base Method. *Journal of Quantitative Linguistics* 7-2, 73 -95.

Holm, Hans J. (2003). The proportionality trap, or: what is wrong with lexicostatistical subgrouping? *Indogermanische Forschungen* 108, 39-47.

Holm, Hans J. (2005). Genealogische Verwandtschaft. In Köhler, R., Altmann, G., and R.G. Piotrowski (Eds), *Quantitative Linguistics*, [HSK-series, vol. 27]. Berlin, New York: DeGruyter, 633-645.

Holm, Hans J. (2007a). Language subgrouping. In Grzybek, P., & R. Köhler (Eds*), Exact Methods in the Study of Language and Text. Festschrift Gabriel Altmann* [Quantitative Linguistics 62]. Berlin: de Gruyter, to appear.

Holm, Hans J. (2007b). The distribution of data in word lists and its impact on the subgrouping of languages. *Proceedings of the 30. GfKl-Meeting 2007*, to appear.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., & J.P. Bollback. (2001). Bayesian Inference of Phylogeny and its Impact on Evolutionary Biology. *Science 294, 2310-2314.*

Jacquesson, François. (2003). Linguistique, génétique et la vitesse d'évolution des langues. *Bulletin de la Société de linguistique de Paris,* t. XCVII, fasc.1, 101-121.

Kendall, David G. (1950). Discussion following Ross, A.S.C., Philological Probability Problems. *Journal of the Royal Statistical Society*, Ser. B 12, 49-50.

Kessler, Brett. (2001). *The significance of word lists*. [CSLI]. Stanford CA: Stanford Univ. Press.

Kim, Junhyong & T. Warnow. (2004). Tutorial on Phylogenetic Tree Estimation, *http://kim .bio.upenn.edu/~jkim/media/ISMBtutorial.pdf.*

Kontzi, R. (1982). *Substrate und Superstrate in den romanischen Sprachen*, Darmstadt: Wissenschaftliche Buchgesellschaft.

Labov, W. (1994). Vol. 1. *Principles of linguistic change.* Oxford: Blackwell.

Lehmann, W.P. (1992). *Historical linguistics: An introduction. London,* New York: Routledge, 3rd ed.

Lehmann, W.P. (1980). Language as a human phenomenon: The importance of history for the understanding of language. *Folia Linguistica Historica*, 1-1, 5-18.

Lohr, M. (2000). New approaches to lexicostatistics and glottochronology. In Renfrew, C., McMahon, A., and L.Trask, *Time depth in historical linguistics,* Vol.1[10]. Oxford: McDonald Institute f. Archaeological Research, 209-222.

McMahon, A., & R. McMahon. (2000). Problems of dating and time depth in linguistics and biology. In Renfrew, C., McMahon, A., and L.Trask, *Time depth in historical linguistics,*Vol.1[4]. Oxford: McDonald Institute f. Archaeological Research, 59-74.

McMahon, A., & R. McMahon. (2003). Finding families: Quantitative methods in language classification. *Transactions of the Philological Society* 101(1), 7-55.

McMahon, A., & R. McMahon. (2005). *Language classification by numbers.* New York: Oxford University Press.

Meier-Brügger, M. (2002). *Indogermanische Sprachwissenschaft*. 8th ed. Berlin: Walter de Gruyter.

Mount, David M. (2001). *Bioinformatics: Sequence and genome analysis.* New York: Cold Spring Harbor Laboratory Press.

Nakhleh, L., Warnow, T., Ringe, D., & S.N. Evans. (2005). A Comparison of Phylogenetic Reconstruction Methods on an IE Dataset. *The Transactions of the Philological Society* 103-2, 171-192.

Pagel, M. (2000). Maximum likelihood models for glottochronology and for reconstructing linguistic phylogenies. In Renfrew, C., McMahon, A., and L.Trask, *Time depth in historical linguistics.* Oxford: McDonald Institute f. Archaeological Research, 189-208.

Pijnenburg, W.J.J. (1983). OIr. Eó, Lat. esox, Basque izoki(n) 'Salmon'. *Orbis* 32, 1-2, 240ff.

Polomé, Edgar C. (1990). The Indo-Europeanization of Northern Europe: The linguistic evidence. *Journal of Indo-European Studies* 18, 3-4, 331-338.

Porzig, Walter. (1954). *Die Gliederung des indogermanischen Sprachgebiets*. Heidelberg: Carl Winter.

Renfrew, C. (1999). Time depth, convergence theory, and innovation in Proto-Indo-European: 'Old Europe' as a PIE linguistic area. *Journal of Indo-European Studies* 27(3,4), 257-293.

Rexová, K., D. Frynta & J. Zrzavý. (2003). Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19, 120-127.

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

Rix, H., Kümmel, M., Zehnder, Th., Lipp, R., and B. Schirmer. (2001). *Lexikon der indogermanischen Verben; die Wurzeln und ihre Primärstammbildungen.* 2nd improved. ed. Wiesbaden: Reichert.

Ruvolo, M. (1987). Reconstructing genetic and linguistic trees: Phenetic and Cladistic approaches. In Hoenigswald, H.M., & L.M. Wiener (Eds), *Biological metaphor and cladistic classification: An interdisciplinary perspective*. Philadelphia, London: Pinter, 193-216.

Schlerath, B. (1992). Rev. of J.P.Mallory (1989) 'In search of the Indo-Europeans' London: Thames & Hudson. *Prähistorische Zeitschrift,* vol 67, 132f.

Schröpfer, Johannes. (1979, cont.). *Wörterbuch der vergleichenden Bezeichnungslehre. Onomasiologie*. Heidelberg: Carl Winter.

Seebold, Elmar. (1981). *Etymologie - eine Einführung am Beispiel der deutschen Sprache*. München: Beck.

Sneath, P.H.A. & R.R. Sokal. (1973). *Numerical taxonomy; the principle and practice of numerical classification.* San Francisco: Freeman.

Sudhaus, W. & K. Rehfeld. (1992). *Einführung in die Phylogenetik und Systematik.* Stuttgart, New York u.a.: G. Fischer.

Swadesh, M. (1952). Lexico-statistic Dating of Prehistoric Ethnic Contacts. *Proceedings of the American Philosophical Society* 96, 452-63.

Swofford, D.L., G.J. Olsen, P.J. Waddell, & D.M. Hillis. (1996). Phylogenetic inference, Chap. 11 in Hillis, D.M., C. Moritz, and B. Mable (Eds), *Molecular Systematics.* 2nd ed., Sunderland, MA: Sinauer, 407-514.

Swofford, D.L. (2002). *PAUP [Phylogenetic Analysis Using Parsimony* (and Other Methods)], Sunderland MA: Sinauer Associates.

Szemerényi, Oswald. (1990). *Einführung in die vergleichende Sprachwissenschaft.* Darmstadt: Wissenschaftliche Buchgesellschaft. 4. Auflage.

Tischler, Johann. (1973). *Glottochronologie und Lexikostatistik.* [Innsbrucker Beiträge zur Sprachwissenschaft 11], Innsbruck.

Tischler, Johann. & B. Ganter. (1997). Review of Dyen/Kruskal/Black: An Indoeuropean Classification. 1992). *Kratylos* 42, 43-50.

Tischler, Johann. (1990). 100 Jahre Kentum-Satem-Theorie. *Indogermanische Forschungen* 95, 63ff.

Thomason, Sarah G. & Terence Kaufmann. (1988). *Language contact, creolization, and genetic linguistics.* Berkeley, University of California Press.

Trask, Larry. (2003). Re: 14.1825, Media: NYT: Keltic Found to Have Ancient Roots. *www.linguistlist.org/issues/14/14-1876.html.*

Viberg, Å. (1983). The Verbs of perception: A typological study. *Linguistics* 21-1(263), 123-62.

Wägele, Johann W. (2001). *Grundlagen der phylogenetischen Systematik*, 2. Ausg. (English version 2005. *Foundations of Phylogenetic Systematics*), München: F. Pfeil.

Wiesemüller, B., Rothe, H. & W. Henke. (2002). *Phylogenetische Systematik: Eine Einführung.* Berlin, New York, u.a.: Springer.

Wilkins, D.P. (1996). Natural tendencies of Semantic Change and the Search for Cognates. In Durie, M. & Ross, M. (Eds). *The comparative method reviewed; regularity and irregularity in language change.* New York: Oxford University Press, 264-304.

Zipf, G.K. (1965). *The psycho-biology of language. An introduction to dynamic philology*. Cambridge MA: MIT Press.