Save the Trees: Why we need tree models in linguistic reconstruction

DRAFT

October 5, 2016

Scepticism against the tree model has a long tradition in historical linguistics. Although scholars have emphasized that the tree model and its longstanding counterpart, the wave theory, are not necessarily incompatible, family trees are unrealistic and should be completely abandoned from historical linguistics has always enjoyed a certain popularity. This scepticism has further increased with recently proposed techniques for data visualization which seem to confirm that we can study language history without trees. In this paper, we show that family trees are not only a logical but also a practical necessity in linguistic reconstruction. While the logical necessity of the tree model follows directly from the basic assumptions underlying linguistic reconstruction, the practical necessity of the tree-model follows from its implications for a realistic modeling of language history, which always needs to involve a before and after of events. In order to save the trees from the critics, we further show that the concrete arguments brought up in favor of anachronistic wave models do not hold. In comparing the phenomenon of *incomplete lineage sorting* in biology with processes in linguistics, we show that data which does not seem to be resolvable in trees may well be explained without turning to diffusion as an explanation. Since the diffusability of features varies, we further show that the failure to find the right tree when relying on easily diffusable features does by no means imply that a tree hypothesis could never be substantiated when using less easily diffusable feature sets. While acknowledging that not all aspects of language history are tree-like, and that integrated models which capture both vertical and lateral language relations may depict language history more realistically, we show that all models which claim that vertical language relations can be completely ignored are essentially wrong: Either they silently still use family trees, or they only provide a static display of data and thus fail to model temporal aspects of language history.

1 Introduction

All languages develop by descent with modification: linguistic material is transferred from generation to generation of speakers, and slight modifications in pronunciation, denotation, and grammar may sum up to changes which are so large that when two or more linguistic varieties have been separated in some way, be it by geographical or political separation of their speakers, they may become mutually incomprehensible. Not all linguistic material is necessarily inherited from the parent generation. Linguistic material may easily be transferred across linguistic boundaries or diffuse across similar speech varieties. This does, however, not change the fact that the primary process by which languages are transmitted is the acquisition of a first language by children. look for ringe quote on this ringe 2002. That largely incomprehensible and different languages may share a common genetic origin was one of the great insights of 19th century linguistics, and even if lateral forces of transmission may drastically change the shape of languages, this does not invalidate the crucial role that genetic inheritance plays in language history. Skepticism against the tree model has a long tradition in historical linguistics. Criticizing the tree model is almost as old as the tree model itself and started not long after the first scholars used evidence drawn from the early version of the comparative method to reconstruct family trees from Indo-European and its subgroups (Čelakovský 1853, Schleicher 1853a, see Geisler & List 2013).

... let's see later how to further expand this ...

2 Dendrophobia and Dendrophilia in the History of Linguistics

In order to get a clearer argument of the major arguments brought up to support or to dismiss the family tree model it is useful to have a closer look at the origins of the tree model and the discussions that it instigated. In the following, we will give a brief overview on the historical background on dendrophilia and dendrophobia in linguistics.

2.1 Dendrophilia

Although not being the first to draw language trees,[1] it was August Schleicher (1821-1866) who made tree-thinking popular in linguistics. In two early papers from 1853 (Schleicher 1853a,b), and numerous studies published thereafter (see, for example, Schleicher 1861, 1863), he propagated that the assumptions about language history could be best 'illustrated by the image of a branching tree' (Schleicher 1853a:787).[2] Note that there was no notable influence by Darwin here. It is more likely that Schleicher was influenced by *stemmatics* (manuscript comparison, see Hoenigswald 1963:8); and even today, historical linguistics has certain features that resemble manuscript comparison much more closely than evolutionary biology. It seems that Schleicher's enthusiasm for the drawing of language trees had quite an impact on Ernst Haeckel (1834-1919, see Sutrop 2012), since – as Schleicher pointed out himself (Schleicher 1863:14) – linguistic trees by then were concrete and not abstract like the one Darwin showed in his Origins (Darwin 1859).

Despite the seemingly radical idea to model language history as process of diversification exclusively via branching and splitting, it is important to note that Schleicher was not a care-

[1] The first trees and networks depicting language development date at least back to the 17th century (for details, see List et al. 2016, Morrison 2016, and Sutrop 2012).

[2] Our translation, original text: '[Diese Annahmen, logisch folgend aus den Ergebnissen der bisherigen Forschung,] lassen sich am besten unter dem Bilde eines sich verästelnden Baumes anschaulich machen'.

less proponent of tree thinking. Judging from his work we find many examples showing that he was aware of potential problems resulting from the tree model. Thus, in his open letter to Haeckel, Schleicher, taking Latin and its descendants as an example, explicitly pointed to problems of language mixing which he compared to plant hybrids in biology, identifying it as a second factor leading to differentiation (Schleicher 1863:18). In his earlier work, he mentioned language contact and borrowing of linguistic features explicitly as a process characteristic for language history (Schleicher 1861:6), emphasizing the importance of distinguishing borrowed from inherited traits in language classification (Schleicher 1848:30). Following up the analogy with species evolution, Schleicher also pointed to the problem of finding sharp borders between languages, dialects, and speech varieties ('Sprache, Dialekt, Mundarten und Untermundarten' in the original), which finds a counter-part in the distinction between species and individuals (Schleicher 1863:21). Especially this last point clearly reflects that Schleicher did not exclusively think that language splits were a product of abrupt separation of speakers, and that he was aware of the idealizing aspect of the *Stammbaum*.

2.2 Dendrophobia

Schleicher's tree-thinking, however, did not last very long in the world of historical linguistics. By the beginning of the 1870s Hugo Schuchardt (1842-1927) and Johannes Schmidt (1843-1901) published critical views, claiming that vertical descent was all what language evolution is about (Schmidt 1872, Schuchardt 1900). While Schmidt remained very vague in his criticism, Schuchardt was concrete and observant in his criticisms, especially pointing to the problem of borrowing between very closely related languages, which might deeply blur the phylogenetic signal:

> We connect the branches and twigs of the family tree with countless horizontal lines and it ceases to be a tree. (Schuchardt 1900:9)[3]

While Schuchardt's observations were based on his deep knowledge of the Romance languages, Schmidt drew his conclusions from a thorough investigation of shared homologous words in the major branches of Indo-European. What he found were patterns of words that were in a strong *patchy distribution* (see List et al. 2014), that is, showing many gaps across the languages, with only a few (if at all) patterns that could be found across all languages. One seemingly surprising fact was, for example, that Greek and Sanskrit shared about 39% of cognates (according to Schmidt's count, see Geisler & List 2013), Greek and Latin shared 53%, but Latin and Sanskrit only 8%. Assuming that Greek and Latin had a common ancestor, Schmidt found it very difficult to explain how the similarities between the two languages with Sanskrit could be so different (Schmidt 1872:24). Furthermore, this pattern of patchy distributions seemed to be repeated in all branches of Indo-European that Schmidt compared in his investigation. Schmidt thus concluded:

[3] Our translation, original text: 'Wir verbinden die Äste und Zweige des Stammbaums durch zahllose horizontale Linien, und er hört auf ein Stammbaum zu sein.'

> No matter how we look at it, as long as we stick to the assumption that today's languages originated from their common proto-language via multiple furcation, we will never be able to explain all facts in a scientifically adequate way. (Schmidt 1872:17)[4]

Schmidt, however, did not stop with this conclusion but proposed another model of language divergence instead of the family tree model:

> I want to replace [the tree] by the image of a wave that spreads out from the center in concentric circles becoming weaker and weaker the farther they get away from the center. (Schmidt 1872:27)[5]

Ever since then, this new model, the so-called *wave theory* (*Wellentheorie* in German) has been vividly discussed in articles and textbooks in in historical linguistics, sometimes being promoted as the missing complement of Schleicher's *Stammbaumtheorie*, sometimes being treated as its more realistic alternative. Although most historical linguists would probably assume that they have a clear understanding of what the wave theory is, it has led to a significant amount of confusion, not only among linguists themselves, but also among those who are not primarily trained in historical linguistics. This confusion is not only reflected in the discussions between dendrophilists and dendrophobists, but also in the various attempts that have been made to visualize the waves. While Schmidt did not give a visualization in his book from 1872, he gave one 3 years later (Schmidt 1875:199), as shown in Figure 1 along with an English translation. It is difficult to interprete this Figure, not only due to the quality, but also due to its structure, which is hard to understand intuitively: It displays languages in a pie-chart-like diagram in a quasi-geographic space. No information regarding ancestral states of the languages is given, and no temporal dynamics are shown. Being quasi-geographic, quasi-quantitative, and quasi-structured, the visualization is hard to understand, and the famous waves themselves are the least thing one thinks about when inspecting it. Schmidt does not seem to ignore that evolution has a time dimension, but he seems to deliberately neglect it when drawing his waves. Other scholars, like Hirt (1905:93), Bloomfield (1973:316), Meillet (1908:134), or Bonfante (1931:174), proposed similar and alternative ways to visualize Schmidt's waves (see Geisler & List 2013). In contrast to the language trees which – after Schleicher's initial rather "realistic" tree drawings – quickly began to be schematized in historical linguistics, the correct way to draw a wave has remained a mysterium up to today.

Visualization problems, however, cannot be taken to discredit a theory, although they may reflect problems of internal coherence. Geisler & List (2013:118-120) distinguish three different kinds of criticisms that have been raised against the family model (and in favor of the wave theory): (1) practicabily problems, (2) plausibility problems, and (3) adequacy problems. Practicability problems refer to the problems in applying the tree model to analyse a given set of languages. In the critics they may be reflected in conflicting evidence, as reflected

[4] Our translation, original text: 'Man mag sich also drehen und wenden wie man will, so lange man an der Anschauung fest hält, dass die in historisches Zeit erscheinenden sprachen durch merfache Gabelungen aus der Ursprache hervorgegangen seien,d.h. so lange man einen stammbaum der indogermanischen Sprachen annimmt, wird man nie dazu gelangen alle die hier in frage stehenden tatsachen wissenschaftlich zu erklären.'

[5] Our translation, original text: 'Ich möchte an seine *[des Baumes]* stelle das bild der welle setzen, welche sich in concentrischen mit der entfernung vom mittelpunkte immer schwächer werdenden ringen ausbreitet.'

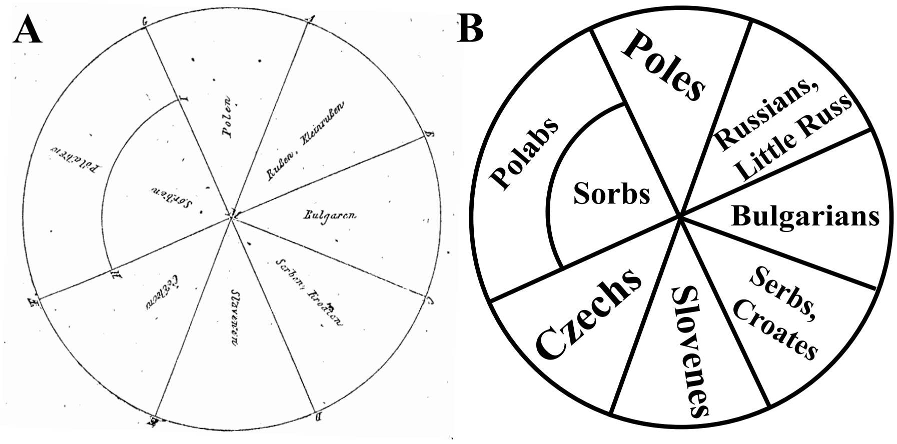


Figure 1: Schmidt's Wave Theory. A: Schmidt's visualization of the Wave Theory from 1875. B: English translation.

in Schmidt's work mentioned above (Schmidt 1872). Plausibility problems refer to the realism of the family tree model and are reflected in obvious simplifications provoked by the tree model. They are reflected in critics emphasizing that languages do not necessarily split abruptly but slowly diverge accompanied by complex waves of diffusion (Schuchardt 1900, Schmidt 1872). Adequacy refers to the purpose of writing language history in historical linguistics. Critics complain that family trees break down all vivid aspects that are substantial for the diversification of a language family to processes of vertical descent.

While all three types of criticism have been brought up against the family tree model, it is clear their theoretical strength differs drastically. Refusing a model for practicality reasons is straightforward, but it cannot be used to prove that a model is wrong or inadequate. Surprisingly, especially the work by Johannes Schmidt reflects a high degree of ignorance regarding the epistemological limits and the temporary status of knowledge in historical linguistics. Being not able to find evidence for a tree in a given dataset is no proof that the family tree model is wrong, in the same way as the inability to distinguish borrowed from inherited traits the further one goes back in time can be considered as proof against the existence of tree-like divergence of languages. Stronger arguments against the family tree model were therefore those that challenged its plausibility, with respect to the presumed split-process by which languages diverge, or its adequacy, with respect to its ability to provide a full picture of language history in all its complexity.

As mentioned earlier, Schleicher was well aware of the most problematic aspects of the tree model, namely the possibility of hybridization and an often gradient as opposed to an abrupt transition underlying language divergence, and he deliberately ignored these aspects in the family tree model, giving a strict preference to divergence and vertical inheritance. Proponents of the wave theory, on the other hand, are much less clear about the different processes they seek to model. Do wave-like processes of language change reflect borrowing among closely related languages, or are they intended to reflect language change in general? While Schuchardt (1900) seems to distinguish the two, pointing to *horizontal lines* ('horizontale Linien') that make a network out of a tree, Schmidt (1872) is much less explicit, although he often invokes the idea of gradual transitions between language borders (Schmidt 1875:200), thus emphasizing the gradualness of diversification rather than the interference of vertical and lateral processes in language change. Given the diversity of opinions and the lack of concreteness, it is difficult to determine a core theory to which scholars refer when mentioning the Wave theory, and while some see the wave theory as the horizontal counterpart of the family tree (Baxter 2006:74), others see the wave theory as a theory explaining linguistic divergence (Campbell 1999:188-191)

2.3 The New Debate on Trees and Waves

Along with the "quantitative turn" in historical linguistics (List 2014:209f), the debate on trees and waves was revived. Had most textbooks treated both models as a complementary view on *external language history*[6] (Lehmann 1992, Anttila 1972) or as treating two completely different aspects of language change (Campbell 1999), more and more linguists now began to discuss the models as opposing perspectives (François 2015). One reason for the revival of the discussion was the prevalence for trees in early phylogenetic studies in historical linguistics (Gray & Atkinson 2003, Atkinson & Gray 2006, Ringe et al. 2002, Pagel 2009). Had both trees and waves been playing a less prominent role for a long time,[7] biological methods for phylogenetic reconstruction applied to large linguistic datasets now offered a quick and much more transparent way to display language data in a tree diagram than the classical method of identifying shared innovations. Only a few years after the first large phylogenetic trees of languages were reconstructed, new visualization techniques for splits networks (most of them based on the NeighborNet algorithm, Bryant & Moulton 2004), as provided by the SplitsTree software package (Huson 1998) offered scholars a fresh view on conflicts in their data which was often propagated as a reconciliation of tree and wave theory (Ben Hamed & Wang 2006, Heggarty et al. 2010, McMahon & McMahon 2005).

3 Theoretical and Practical Necessity of the Tree Model

as in the longer abstract above, but of course, more verbose, with more references, basically, this is, what was before labelled as:

- The tree model as a logical consequence of the comparative method

Part written by M + G

[6] External language history is here used in the sense of Gabelentz (1891) who distinguishes it from internal language history pointing to different stages of one and the same language.

[7] Even Morris Swadesh never used his lexicostatistic method to produce family trees. Instead, he published a map on "interrelationships of American Indian languages" that comes closer to an interpretation in terms of the wave theory (Swadesh 1959:23).

4 The Advantages of the Tree Model

Here, it would be the parts mentioned not in he abstract, but in the first draft, with arguments by G

5 Saving the Trees from the Critics

Given the logical necessity to allow for divergence, a specific part of language history can be modeled with help of a tree if specific processes like recombination (hybridization, creolization) can be excluded. That such a tree model does not necessarily represent all aspects of language history is obvious, and even the strongest tree proponents would not deny it. Whether the amount of inheritance versus borrowing in language history is as low as it was supposed for biology, where tree critics have labeled the tree of life as the "tree of one percent" (Dagan & Martin 2006) is an interesting question worth being pursued further. Given that we know that language varieties can diverge to such an extent that they loose mutual intelligibility, however, necessitates a model for language history which handles divergence and splits of lineages. How these splits proceed in the end, whether they are best viewed as multifurcations after the split of a larger dialect continuum in several parts, or as bifurcations, depends on our insights into the language family under investigation and into the processes of external language change in general.

5.1 Not Seeing the Tree for the Forest

When scholars point out that a given datasets lacks tree-like signal, or that the tree-like signal for the subgrouping of a given language family is not strong, they often take this as direct evidence for large-scale language contact or linkage scenarios (Ross 1988). This, however, is by no means the only explanation for reticulations in datasets, and many other reasons why a given data selection may fail to reveal a tree (see the general overview in Morrison 2011:44-66). The most obvious and in cases of large dataset most frequent reason are erroneous codings which occur especially in those cases where the data has not been thoroughly checked by the experts in the field (Geisler & List 2010), or where automatic analyses have introduced a strong bias. Another obvious reason is the selection of the data. Commonalities in sound change patterns and grammatical features, for example, often do not represent true shared innovations but independent development, and especially for sound changes it is often very hard to distinguish between synapomorphy and homplasy (Chacon 2015:182f), which is exacerbated by the fact that the majority of sound change patterns are extremely common, while rare sound changes are often very difficult to prove. Apart from borrowing, dialect differentiation, data coding, and homoplasy, another often overlooked cause of reticulations in the data is the process of *incomplete lineage sorting* (Galtier & Daubin 2008). Incomplete lineage sorting is a well-known process in biology, during which polymorphisms in the ancestral lineages are inherited by the descendent languages when rapid divergence occurs (Rogers & Gibbs 2014). Incomplete lineage sorting can explain why 30% of the genes in a Gorilla's genome are more similar to the human genome or to the Chimpanzee genome although human and chimpanzee are the closest relatives (Scally et al. 2012). In a recent study, List et al. (2016) proposed that incomplete lineage sorting may likewise occur in language history, given the multiple sources of polymorphisms in language change, ranging from near synonymy of lexical items via suppletive paradigms to word derivation. Apart from these language-internal factors of polymorphisms which may be inherited across lineages, before they are later randomly resolved, an further factor not mentioned by List et al. (2016) is variation in the population of speakers, or sociolinguistic variation, which, in contrast to biology, where every organism has only one gene for each function, may even occur in one and the same speaker. The process of incomplete lineage sorting is further illustrated in Figure 2, where the two aspects, namely sociolinguistic variation, and language-internal variation are contrasted. Note that in neither of the cases we even need to invoke neither strong language contact nor situations of large scale diffusion in dialect networks. Both patterns are perfectly compatible with a "social split" situation as invoked by François (2015), although they are based on fully resolved bifurcating trees. This shows that supposed reticulations in the data, or lack of tree-like signal in the data do not necessarily prove the absence of tree-like patterns of divergence. They rather expose the weakness of our methods to find the tree in the forest of individual histories of linguistic traits.

5.2 Diffusability of Features

Here, G formulates the optimistic message that we can RANK our features and do not need to be as pessimistic as the glottometric-people

5.3 Competing forms as a cause of reticulation

François (2015:178) puts much value on *lexically-specific sound changes*, arguing that they are "strongly indicative of genealogy, because they are unlikely to diffuse across separate languages". Out of his 474 innovations, 116 (24%) belong to this type. In view of the low diffusibility of such innovations,[8] overlapping isoglosses constitute in his view a major problem for the tree model. In this section, we argue that regardless of whether lexically-specific sound changes have more difficulty to cross language boundaries than other types of innovations, overlapping innovations can nevertheless be accounted for by assuming the existence of competing variants in the proto-language.

Languages are never completely uniform, and fieldwork linguists working on unwritten languages commonly notice that even siblings may present significant differences in the pronunciation of certain words or even morphological paradigms (see for instance Genetti 2007:29-30).

While some innovations can spread quickly to the entire community (or at least to all members of a specific age-group), in other cases it is possible for two competing forms (innovative vs archaic) to remain used in the same speech community for a considerable period of time. This is observed in particular with sporadic changes, such as irregular metatheses / dissimilation / assimilation, or item-specific analogy.

When language differentiation occurs while forms are still competing, daughter languages can inherit the competing forms; then the innovative form may eventually prevail or disappear

[8] This assertion remains to be demonstrated, but we accept it for the sake of argument.

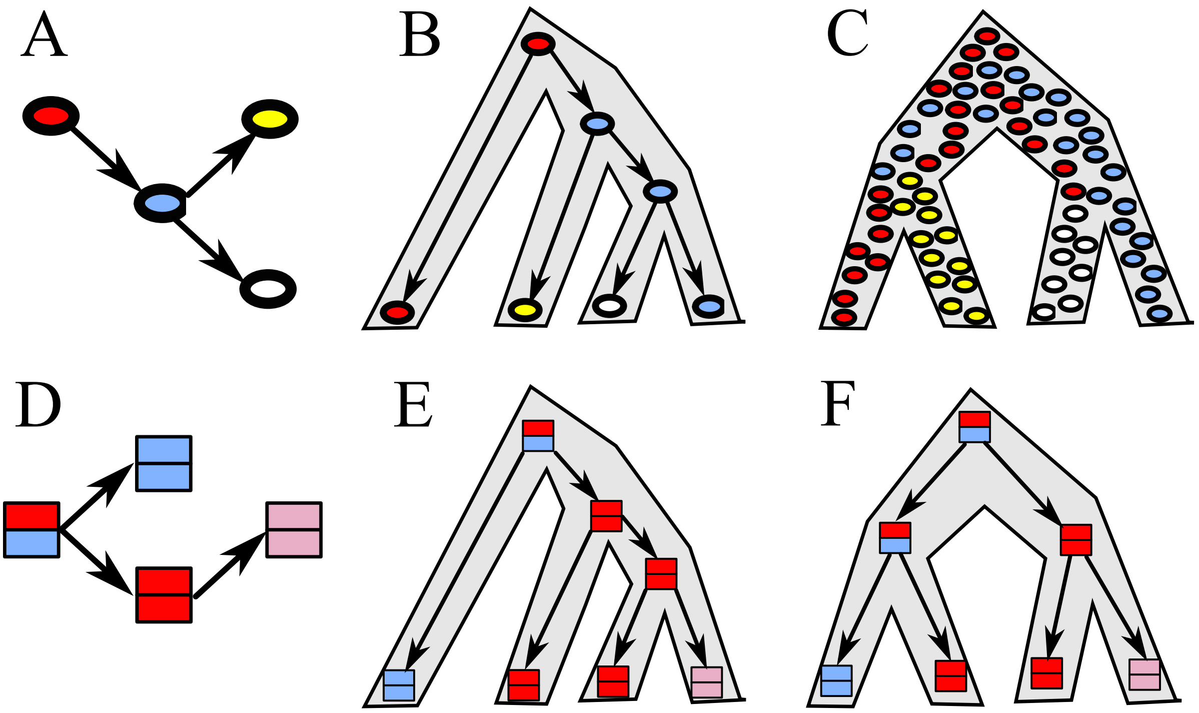


Figure 2: Incomplete lineage sorting due to sociolinguistic (A-C) and linguistic variation (D-F) and its impact on phylogenetic reconstruction and genetic subgrouping. A shows a pattern of know directional evolution of a character (e.g., a sound change pattern), and B shows one of the most parsimonious trees resulting from the pattern. C shows an alternative pattern by assuming that the blue character already evolved in the ancestral language where it was used as a variant along with the original red character. Since the variation already occurred at the time of the ancestral languages, it was inherited in the two descendant languages from which the character further developed. As a result, another tree topology can be reconstructed. D gives an example for a process of paradigm leveling, and E and F show two possible equally parsimonious scenarios invoking different tree topologies each.

in a non-predictable way in each daughter language. If such situation occurs, the distribution of the innovation will not be relatable to the phylogenetic tree.

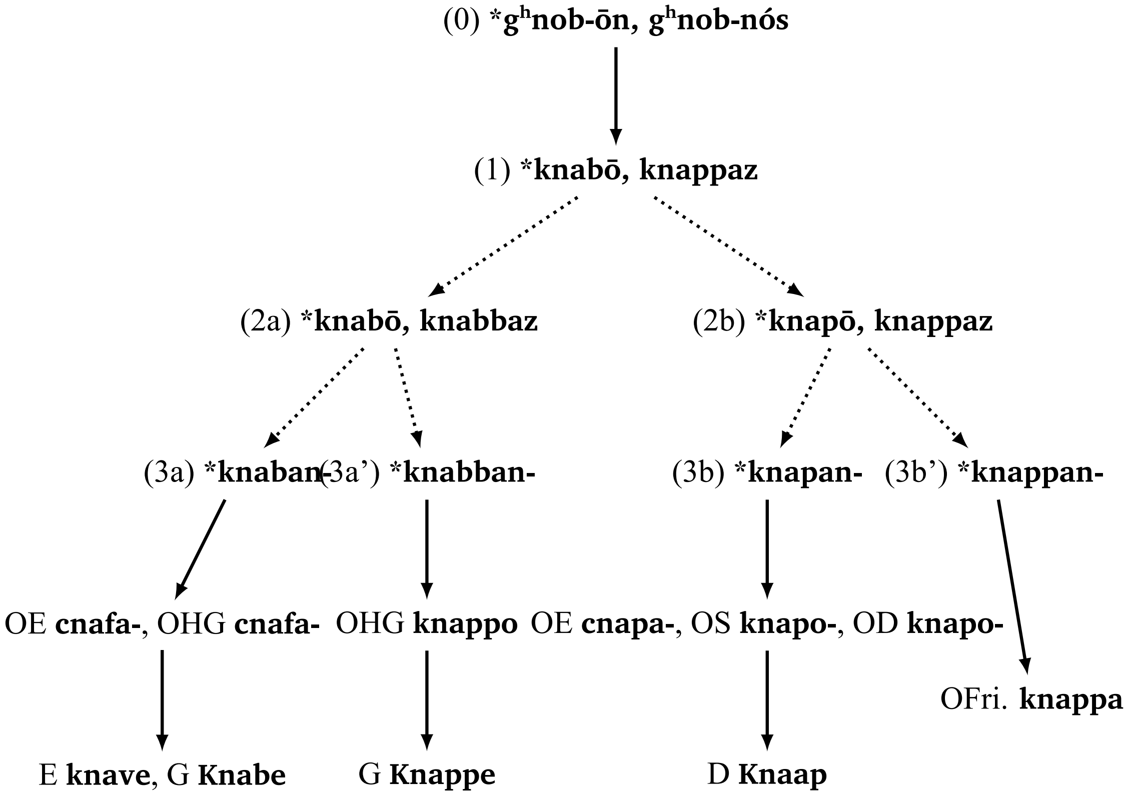
This phenomenon is better illustrated by analogical levelling rather than by sporadic sound changes, as in the case of the former the variation comes from well-understood morphological alternations that have been generalized in different ways in different language varieties, though the same account would be valid of the sporadic changes.

To illustrate how alternation and variation in the proto-language can blur the phylogeny, we take the example of the reflexes of proto-Germanic ***knabō, knappaz** 'boy' (Figure 3),[9] an n-stem noun whose reflexes in the modern and ancient languages are particularly complex (data from Kroonen 2011:71,128, Kroonen 2013:294).

[9] The reflexes of this proto-form have developped distinct meanings in the attested languages, including 'squire', but this aspect is not considered here.

From the attested ancient and modern forms (if the known sound laws are applied backwards), no less than four protoforms have to be postulated: ***knaban-**, ***knapan-**, ***knabban-** and ***knappan-**. Some languages have more than one reflex of this etymon (with diverging specialized meanings), and their distribution does not fit any accepted classification of the Germanic languages: for instance, while nearly all Germanicists agree on the existence of an Anglo-Frisian 'Ingvaeonic' branch, we see that English sides with either German (in have a reflex of ***knaban-**) or with Dutch (the Old English reflex of ***knapan-**, lost in modern English) rather than Frisian.

Figure 3: Several layers of variation: the etymon ***knabō, knappaz** 'boy' in Germanic



Unlike most other language families, the detailed knowledge that has accumulated on the history of Germanic languages allows to go further than stating the presence of irregular correspondences: it is possible to account for them with a detailed model. It is now near-universally accepted that doublets such as those are due to the effect of Kluge's law on the endings of n-stem nouns in pre-proto-Germanic (stage 0, Kluge 1884, Kroonen 2011).

The paradigm of the noun 'boy' (and all nouns of the same type) in proto-Germanic (stage 1) had an alternation between ***-b-** and ***-pp-**. This complex alternation was variously levelled as ***-b-/-bb-** or ***-p-/-pp-** by stage 2; note that within a single language, not all items belonging to this declension class underwent levelling in the same way, and that some languages even have competing innovative (OE **cnapa** from ***knapan-**) and archaic (OE **cnafa** from ***knaban-**)

forms for the same etymon (in this particular case, no that only the archaic form has been preserved, with a different meaning, in modern English **knave**).

After simplification of the ***-b-**/***-pp-**, all languages underwent a second wave of analogy, generalizing either the stem of the nominative (archaic ***knaban-** or innovative ***knapan-**) or that of the genitive (archaic ***knappan-** or innovative ***knabban-**), resulting in the four variants attested throughout Germanic languages.

We do not deny the potential value of item-specific changes of this type as evidence for studying phylogeny. However, it is obvious that isoglosses based on item-specific analogical levelling and sporadic sound change will overlap with each other, since competing forms can be maintained within the same language variety.

5.4 Limitations of Sound Correspondences to Identify Lexical Innovations

In order to identify inherited lexical innovations and distinguish them from recent borrowings, François (2015:176-8) uses a fairly uncontroversial criterion: etyma whose reflexes follow regular sound correspondences are considered to be inherited. Thus, whenever a common proto-form can be postulated for a particular set of words across several languages (which can thus be derived from this proto-form by the mechanical application of regular sound changes), it is considered in this model to be part of the inherited vocabulary, and can be used, if applicable, as a common innovation.

François' approach however neglects an important factor: while regular sound correspondences is a necessary condition for analyzing forms in related languages as cognates, i.e. originating from the same etymon in their common ancestor,[10] it is not a **sufficient** condition due to the existence of **undetectable borrowings** and **nativized loanwords**.

5.4.1 Undetectable borrowings

Sound changes are not always informative enough to allow the researcher to discriminate between inherited word and borrowing. When a form contains phonemes that remained unchanged, or nearly unchanged, from the proto-language in all daughter languages (because no sound change, or only trivial changes, affected them), there is no way to know whether it was inherited from the proto-language or whether it was borrowed at a later stage.

This type of situation is by no means exceptional, and can be found in various language families. We present here three examples of borrowings undetectable by phonology alone: 'aluminum' in Tibetan languages, 'pig' in some Algonquian languages, and 'palace' in Semitic.

Amdo Tibetan **hajaŋ** 'aluminum' and Lhasa **hájã** 'aluminum' look like they regularly originate from a Common Tibetan form ***ha.jaŋ**.[11] This is of course impossible for obvious his-

[10] Note however that cognacy is a more complex concept that is usually believed (List 2016), and that even forms originating from exactly the same etymon in the proto-language may present irregular correspondences due to analogy.

[11] In Amdo Tibetan, Common Tibetan **h-**, **j-**, **-a** and **-aŋ** remain unchanged (Gong 2016). In Lhasa Tibetan, two sound changes relevant to this form occurred: a phonological high tone developped with the initial **h-**, and **-aŋ** became nasalized **ã**.

torical reasons, as aluminum came into use in Tibetan areas in the twentieth century, at a time when Amdo Tibetan and Lhasa Tibetan were already completely unintelligible. This word is generally explained (Gong Xun, p.c.) as an abbreviated form of **ha.tɕaŋ jaŋ.po** ‘very light’, but this etymology is not transparent to native speakers of either Amdo or Lhasa Tibetan. This word has been coined only once,[12] and was then borrowed into other Tibetan languages[13] and neighboring minority languages under Tibetan influence (as for instance Japhug **χajaŋ** ‘aluminum’).

In this case a phonetic borrowing from Amdo **hajaŋ** could only yield Lhasa Lhasa **hájã**, since **h-** only occurs in high tone in Lhasa, and since final **-ŋ** has been transphonologized as vowel nasality.[14]

Several Algonquian languages, share a word for ‘pig’ (Fox **koohkooša**, Miami **koohkooša** and Cree **kôhkôs**) ultimately of Dutch origin (Goddard 1974, Costa 2013). Hockett (1957:266) pointed out that these forms must be considered to be loanwords ‘because of the clearly post-Columbian meaning; but if we did not have the extralinguistic information the agreement in shape (apart from M[enominee]) would lead us to reconstruct a [Proto-Central-Algonquian] prototype.’ The forms from these three languages could be regularly derived from Proto-Algonquian ***koohkooša**, a reconstruction identical to the attested Fox and Miami forms.

Semitic languages abound in common vocabulary which presents the same correspondences as inherited vocabulary, but which was diffused after the breakup of the family. For instance, from Biblical Hebrew **hêkāl** ‘palace’ and Arabic **haykal** ‘palace’, it would be possible to reconstruct a Proto-Semitic etymon ***haykal(u)**; it is however well-known that these words originate from Sumerian **é.gal** ‘palace’, probably through Akkadian **ekallum**, and that borrowing from Akkadian took place at a time when the ancestors of Hebrew and Arabic respectively were already distinct languages.

Undetectable borrowings is also a pervasive phenomenon in Pama-Nyungan, where with a few exception such as the Arandic and Paman groups, most languages present too few phonological innovations to allow easy discrimination for loanwords from cognates (Koch 2004:46).

The same situation can be observed even if latter sound changes apply to both borrowings and inherited words. Whenever borrowing takes place after the break-up of two languages, but before any diagnostic sound change occurred in either the donor or the receiver language, phonology alone is not a sufficient criterion to distinguish between inherited words and loanwords.

A classical case is that of Persian borrowings in Armenian. As Hübschmann (1897:16-17) put it, ‘in isolated cases, the Iranian and the genuine Armenian forms match each other phonetically, and the question whether borrowing [or common inheritance] has to be assumed

[12] We are not aware of a detailed historical research on the history of this particular word, but in any case it matters little for our demonstration whether it was first coined in Central Tibetan or in Amdo.

[13] In some Tibetan languages such as Cone **hæ̀jãː**, Jacques (2014:306), there is clear evidence that the word is borrowed from Amdo Tibetan and is not native (otherwise †**hæ̀jaː** would have been expected).

[14] Likewise, in the case of borrowing from Lhasa into Amdo, the rhyme **-aŋ** would be the only reasonable match for Lhasa **-ã**.

must be decided from a non-linguistic point of view.'[15] Table 1 presents a non-exhaustive list of such words.

Table 1: Armenian words which cannot be conclusively demonstrated to be either borrowings from Iranian or inherited words from a phonetic point of view

Armenian	Meaning	Indo-Iranian	Reference
naw	boat	Skt. **nau-**	Hübschmann (1897:16-17;201), Martirosyan (2010:466;715)
mēg	mist	Skt. **megha-**, Avestan **maēɣa-**	Hübschmann (1897:474), Martirosyan (2010:466;715)
mēz	urine	Skt. **meha-**	Hübschmann (1897:474), Martirosyan (2010:466;715)
sar	head	Skt. **śiras-** Y.Avestan **sarah-**	Hübschmann (1897:236;489), Martirosyan (2010:571)
ayrem	burn	Skt. **edh-**	Hübschmann (1897:418), Martzloff (2016:145)

The Armenian case shows that undetectable loans are not restricted to cases like those studied above, when a particular word only contains segments which have not been affected by sound changes from the proto-language to all its daughter languages. Undetectable loans are possible when a particular word is borrowed before any sound change which could affect its phonetic material occurred in either the giver or recipient language, even if numerous sound changes occurred *after* borrowing took place. It is possible that post-borrowing sound changes may even remove phonetic clues which could have allowed to distinguish between loanwords and inherited words.

We have shown clear evidence that undetectable borrowings can occur even when two language varieties are mutually unintelligible. Neglecting the distinction between inherited words and undetectable borrowings, as in François' model, amounts to losing crucial historical information.

5.4.2 Nativization of loanwords

In the previous section, we have discussed cases when borrowing took place before diagnostic sound changes, thus making it impossible to effectively use sound changes to distinguish between loanwords and inherited words. There is however evidence that even when diagnostic sound changes exist, they may not always be an absolutely reliable criterion.

When a particular language contains a sizeable layer of borrowings from another language, bilingual speakers can develop a intuition of the phonological correspondences between the

[15] 'In einzelnen Fällen kann allerdings das persische und echt armenische Wort sich lautlich decken und die Frage, ob Entlehnung anzunehmen ist oder nicht, muss dann nach andern als sprachlichen Gesichtspunkten entschieden werden.'

two languages, and apply these correspondences to newly borrowed words, a phenomenon known as loan nativization.

The best documented case of loan nativization is that between Saami and Finnish (the following discussion is based on Aikio 2006). Finnish and Saami are only remotely related within the Finno-Ugric branch of Uralic, but Saami has borrowed a considerable quantity of vocabulary from Finnish, some at a stage before most characteristic sound changes had taken place, other more recently. Table 2 presents examples of cognates between Finnish and Saami illustrating some recurrent vowel and consonant correspondences.

Table 2: Examples of of sound correspondences in inherited words between Finnish and Saami (data from Aikio 2006:27)

Finnish	Saami	Proto-Finno-Ugric	Meaning
käsi	**giehta**	***käti**	‘hand’
nimi	**namma**	***nimi**	‘name’
kala	**guolli**	***kala**	‘fish’
muna	**monni**	***muna**	‘egg’

The correspondence of final **-a** to **-i** and final **-i** to **-a** in disyllabic words found in the native vocabulary, as illustrated by the data in Table 2, is also observed in Saami words borrowed from Finnish, including recent borrowings, such as **mearka** from **merkki** ‘sign, mark’ and **báhppa** from **pappi** ‘priest’ (from Russian **поп**, itself of Greek origin), even though the sound change from proto-Uralic to Saami leading to the correspondence **-a** : **-i** had already taken place at the time of contact. These correspondences are pervasive even in the most recent borrowings, to the extent that according to Aikio (2006:36) ‘examples of phonetically unmarked substitutions of the type F[innish] **-i** > Saa[mi] **-i** and F[innish] **-a** > Saa[mi] **-a** are practically nonexistent, young borrowings included.’

In cases such as **báhppa** ‘priest’, the vowel correspondence in the first syllable **á** : **a** betrays its origin as a loanword, as the expected correspondence for a native word would be **uo** : **a** as in the word ‘fish’ in Table 2 (Aikio 2006:35 notes that this correspondence is never found in borrowed words).

However, there are cases where recent loanwords from Finnish in Saami present correspondences indistinguishable from those of the inherited lexicon, as **barta** ‘cabin’ from Finnish **pirtti**, itself from dialectal Russian **перть** ‘a type of cabin’, which show the same **CiCi** : **CaCa** vowel correspondence as the word ‘name’ in table 2. Here again, the foreign origin of this word is a clear indication that **barta** ‘cabin’ cannot have undergone the series of regular sound changes leading from proto-Finno-Ugric ***CiCi** to Saami **CaCa**, and that instead the common vowel correspondence **CiCi** : **CaCa** was applied to Finnish **pirtti**.

Loan nativization can also occur between genetically unrelated languages. A clear example is provided by the case of Basque and Spanish (Trask 2000:53-54, Aikio 2006:21-3).

A recurrent correspondence between Spanish and Basque is word-final **-ón** to **-oi**. Early Romance ***-one** (from Latin **-onem**) yields Spanish **-ón**. In Early Romance borrowings into Basque, however, this ending undergoes the regular loss of intervocalic ***-n-** (a Basque-internal sound change), and yields ***-one** → ***-oe** → **-oi**. An example of this correspondence is provided by Spanish **razón** and Basque **arrazoi** 'reason' both from Early Romance ***ratsone** (from the Latin accusative form ← **ratiōnem**).

This common correspondence have however been recently applied to recent borrowings from Spanish such as **kamioi** 'truck' and **abioi** 'plane' (from **camión** and **avión**). This adaptation has no phonetical motivation, since word-final **-on** is attested in Basque, and can only be accounted for as overapplication of the **-oi** : **-ón** correspondence.

5.4.3 Implications for the interpretation of François' data

In his database of 474 innovations, François (2015:177) counts 236 lexical replacements and 116 lexically-specific sound changes (in other words, lexical replacements where the innovative form is related to the archaic form by some unpredictable phonetic change). A group of languages are considered to share these innovations whenever they present reflexes of these etyma following regular correspondences.

Thus, 352 out of 474 innovations (74%) are potential cases of either undetectable or nativized borrowings. XXX

6 Conclusion

following passage from long abstract probably useful While acknowledging that not all aspects of language history are tree-like, and that integrated models which capture both vertical and lateral language relations may depict language history more realistically, we show that all models which claim that vertical language relations can be completely ignored are essentially wrong. Either they silently still use family trees, or they only provide a static display of data and thus fail to model temporal aspects of language history.

References

Aikio, Ante. 2006. Etymological Nativization of Loanwords: a Case Study of Saami and Finnish. In Ida Toivonen & Diane Nelson (eds.), *Saami Linguistics*, 17–52. Amsterdam: Benjamins.

Anttila, Raimo. 1972. *An introduction to historical and comparative linguistics*. New York: Macmillan.

Atkinson, Quentin D. & Russell D. Gray. 2006. How old is the Indo-European language family? Illumination or more moths to the flame? In Peter Forster & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 91–109. Cambridge and Oxford and Oakville: McDonald Institute for Archaeological Research.

Baxter, W. H. 2006. Mandarin dialect phylogeny. *Cahiers de Linguistique – Asie Orientale* 35(1). 71–114.

Ben Hamed, Mahe & Feng Wang. 2006. Stuck in the forest: Trees, networks and chinese dialects. *Diachronica* 23. 29–60.

Bloomfield, Leonard. 1973. *Language*. London: Allen & Unwin.

Bonfante, G. 1931. I dialetti indoeuropei. *Annali del R. Istituto Orientale di Napoli* 4. 69–185.

Bryant, David & V. Moulton. 2004. Neighbor-net. *Molecular Biology and Evolution* 21(2). 255–265.

Campbell, Lyle. 1999. *Historical linguistics. an introduction*. Edinburgh: Edinburgh Univ. Press 2nd edn.

Chacon, Thiago Costa. 2015. The reconstruction of laryngealization in proto-tukanoan. In Matthew Coler (ed.), *Laryngeal features in the languages of the americas*, 258–284. Leiden: Brill.

Costa, David J. 2013. Borrowing in Southern Great Lakes Algonquian and the History of Potawatomi. *Anthropological Linguistics* 55(3). 195–233.

Dagan, Tal & William Martin. 2006. The tree of one percent. *Genome Biology* 7(118). 1–7.

Darwin, Charles. 1859. *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. London: John Murray. Electronic resource. Online available under: http://www.nla.gov.au/apps/cdview/nla.gen-vn4591931.

François, Alexandre. 2015. Trees, waves and linkages: models of language diversification. In Claire Bowern & Bethwyn Evans (eds.), *The routledge handbook of historical linguistics*, 161–189. Routledge.

Gabelentz, Hans Georg C. 1891. *Die sprachwissenschaft*. Leipzig: T. O. Weigel.

Galtier, Nicolas & Vincent Daubin. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 363(1512). 4023–4029. doi:10.1098/rstb.2008.0144. http://rstb.royalsocietypublishing.org/content/363/1512/4023.

Geisler, H. & J.-M. List. 2013. Do languages grow on trees? the tree metaphor in the history of linguistics. In Heiner Fangerau, Hans Geisler, Thorsten Halling & William Martin (eds.), *Classification and evolution in biology, linguistics and the history of science. concepts – methods – visualization*, 111–124. Stuttgart: Franz Steiner Verlag.

Geisler, Hans & Johann-Mattis List. 2010. Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics. In Heinrich Hettrich (ed.), *Die ausbreitung des indogermanischen. thesen aus sprachwissenschaft, archäologie und genetik*, Wiesbaden: Reichert. Document has been submitted in 2010 and is still waiting for publication.

Genetti, Carol. 2007. *A Grammar of Dolakha Newar*. Berlin: Mouton de Gruyter.

Goddard, Ives. 1974. Dutch Loanwords in Delaware. In Herbert C. Kraft (ed.), *A Delaware Indian Symposium*, 153–60. Harrisburg: Pennsylvania Historical and Museum Commission.

Gong, Xun. 2016. A phonological history of Amdo Tibetan rhymes. *Bulletin of the School of African and Oriental Studies* 79(2). 347–374.

Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965). 435–439.

Heggarty, P., W. Maguire & A. McMahon. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 365(1559). 3829–3843.

Hirt, Herman. 1905. *Die Indogermanen.*, vol. 1. Strassburg: Trübner.

Hockett, Charles F. 1957. Central Algonquian Vocabulary: Stems in k-. *International Journal of American Linguistics* 23(4). 247–268.

Hoenigswald, Henry M. 1963. On the history of the comparative method. *Anthropological Linguistics* 5(1). pp. 1–11.

Huson, Daniel H. 1998. Splitstree: analyzing and visualizing evolutionary data. *Bioinformatics* 14(1). 68–73.

Hübschmann, Heinrich. 1897. *Armenische Grammatik. 1. Theil: Armenische Etymologie*. Leipzig: Breitkopf und Härtel.

Jacques, Guillaume. 2014. Cone. In Jackson T.-S. Sun (ed.), *Phonological Profiles of Little-Studied Tibetic Varieties*, 265–371. Taipei: Academia Sinica.

Kluge, F. 1884. Die germanische Consonantendehnung. *Beitrage der deutschen Sprache und Literatur* 9. 149–186.

Koch, Harold. 2004. A methodological history of Australian linguistic classification. In Claire Bowern & Harold Koch (eds.), *Australian Languages: Classification and the Comparative Method*, 17–60. Amsterdam: Benjamins.

Kroonen, Guus. 2011. *The Proto-Germanic n-stems, A study in diachronic morphophonology*. Amsterdam: Rodopi.

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

Lehmann, Winfred Philipp. 1992. *Historical linguistics*. London: Routledge 3rd edn.

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

List, Johann-Mattis. 2016. Beyond cognacy: historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1(2). 119–136. doi:10.1093/jole/lzw006.

List, Johann-Mattis, Shijulal Nelson-Sathi, Hans Geisler & William Martin. 2014. Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *Bioessays* 36(2). 141–150.

List, Johann-Mattis, Jananan Sylvestre Pathmanathan, Philippe Lopez & Eric Bapteste. 2016. Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics. *Biology Direct* 11(39). 1–17.

Martirosyan, Hrach K. 2010. *Etymological dictionary of the Armenian inherited lexicon*. Leiden: Brill.

Martzloff, Vincent. 2016. Arménien geri 'captif' : étymologie et phraséologie à la lumière d'un parallèle en albanien du Caucase. *Wékwos* 2. 109–179.

McMahon, April & Robert McMahon. 2005. *Language classification by numbers*. Oxford: Oxford University Press.

Meillet, Antoine. 1908. *Les dialectes Indo-Européens*. Paris: Librairie Ancienne Honoré Champion.

Morrison, D. A. 2011. *An introduction to phylogenetic networks*. Uppsala: RJR Productions.

Morrison, David A. 2016. Genealogies: Pedigrees and phylogenies are reticulating networks not just divergent trees. *Evolutionary Biology* doi:10.1007/s11692-016-9376-5. Published online before print.

Pagel, Mark. 2009. Human language as a culturally transmitted replicator. *Nature Reviews. Genetics* 10. 405–415.

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

Rogers, J. & R. A. Gibbs. 2014. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat. Rev. Genet.* 15(5). 347–359.

Ross, Malcom D. 1988. *Proto-Oceanic and the Aaustronesian languages of Western Melanesia*. Canberra: Pacific Linguistics.

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques-Bonet, S. McCarthy, S. H. Montgomery, P. C. Schwalie, Y. A. Tang, M. C. Ward, Y. Xue, B. Yngvadottir, C. Alkan, L. N. Andersen, Q. Ayub, E. V. Ball, K. Beal, B. J. Bradley, Y. Chen, C. M. Clee, S. Fitzgerald, T. A. Graves, Y. Gu, P. Heath, A. Heger, E. Karakoc, A. Kolb-Kokocinski, G. K. Laird, G. Lunter, S. Meader, M. Mort, J. C. Mullikin, K. Munch, T. D. O'Connor, A. D. Phillips, J. Prado-Martinez, A. S. Rogers, S. Sajjadian, D. Schmidt, K. Shaw, J. T. Simpson, P. D. Stenson, D. J. Turner, L. Vigilant, A. J. Vilella, W. Whitener, B. Zhu, D. N. Cooper, P. de Jong, E. T. Dermitzakis, E. E. Eichler, P. Flicek, N. Goldman, N. I. Mundy, Z. Ning, D. T. Odom, C. P. Ponting, M. A. Quail, O. A. Ryder, S. M. Searle, W. C. Warren, R. K. Wilson, M. H.

Schierup, J. Rogers, C. Tyler-Smith & R. Durbin. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388). 169–175.

Schleicher, August. 1848. *Zur vergleichenden Sprachengeschichte*. Bonn: König.

Schleicher, August. 1853a. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur* 3. 786–787.

Schleicher, August. 1853b. O jazyku litevském, zvlástě ohledem na slovanský. Čteno v posezení sekcí filologické král. České Společnosti Nauk dne 6. června 1853. *Časopis Čsekého Museum* 27. 320–334.

Schleicher, August. 1861. *Kurzer abriss einer lautlehre der indogermanischen ursprache*, vol. 1. Weimar: Böhlau.

Schleicher, August. 1863. *Die darwinsche theorie und die sprachwissenschaft*. Weimar: Hermann Böhlau.

Schmidt, Johannes. 1872. *Die verwantschaftsverhältnisse der indogermanischen sprachen*. Hermann Böhlau.

Schmidt, Johannes. 1875. *Zur geschichte des indogermanischen vocalismus. zweite abteilung*. Weimar: Hermann Böhlau.

Schuchardt, Hugo. 1900. *Über die Klassifikation der romanischen Mundarten. Probe-Vorlesung, gehalten zu Leipzig am 30. April 1870*. Graz.

Sutrop, Urmas. 2012. Estonian traces in the tree of life concept and in the language family tree theory. *Journal of Estonian and Finno-Ugric Lingusitics* 3. 297–326.

Swadesh, Morris. 1959. The mesh principle in comparative linguistics. *Anthropological Linguistics* 1(2). 7–14.

Trask, Larry. 2000. Some Issues in Relative Chronology. In Colin Renfrew, April McMahon & Larry Trask (eds.), *Time Depth in Historical Linguistics, vol. 1*, 45–58. Cambridge, UK: The McDonald Institute for Archaeological Research.

Čelakovský, F. L. 1853. *Čtení o srovnavací mluvnici slovanské*. Prague: V komisí u F. Řivnáče.