*Journal of Language Evolution*, 2016, 119–136
doi: 10.1093/jole/lzw006
Advance Access Publication Date: 28 June 2016
Research article

OXFORD

# Beyond cognacy: historical relations between words and their implication for phylogenetic reconstruction

**Johann-Mattis List***

Centre des Recherches Linguistiques sur l'Asie Orientale, École des Hautes Études en Sciences Sociales, 2 rue de Lille, Paris 75007, France, and Team Adaptation, Integration, Reticulation, Evolution, Université Pierre et Marie Curie, 9 quai St Bernard, Paris 75005, France

*Corresponding author: mattis.list@lingpy.org

## Abstract

This article investigates the terminology and the processes underlying the fundamental historical relations between words in linguistics (*cognacy*) and genes in biology (*homology*). The comparison between linguistics and biology shows that there are major inconsistencies in the analogies drawn between the two research fields and the models applied in phylogenetic reconstruction in linguistics. Cognacy between words is treated as a binary relation which is either present or not. Words, however, can exhibit different degrees of cognacy which go beyond the distinction between orthologous and paralogous genes in biology. The complex nature of cognacy has strong implications for the models used for phylogenetic reconstruction. Instead of modeling lexical evolution as a process of cognate gain and cognate loss, we need to go beyond the cognate relation and develop models which take the *degrees of cognacy* into account. This opts for the use of evolutionary models which handle multistate characters and allow to define potentially asymmetrical transition tendencies among the character states instead of time-reversible binary state models in phylogenetic approaches. The benefit of multistate models with asymmetric transition tendencies is demonstrated by testing how well different models of lexical change perform in semantic reconstruction on a lexicostatistical dataset of 23 Chinese dialects in a parsimony framework. The results show that the improved models largely outperform the popular gain–loss models. This suggests that improved models of lexical change may have strong consequences for phylogenetic approaches in linguistics.

## 1. Introduction

Evolutionary biology and historical linguistics both deal with the evolution of objects. Evolutionary biology investigates the evolution of species, morphological characters, and genes, and historical linguistics investigates the evolution of language varieties, grammatical features, and words. In both disciplines, *historical relations* are an important way to describe the consequences of evolutionary processes. Historical relations are defined for evolving objects which share a common history. The most general historical relation is the relation of *common descent*. This relation can hold both for lineages and for their characteristics. If the relation concerns the latter, biologists call it *homology*. In linguistics, this relation is often compared with the relation of *cognacy*. In contrast to historical relations, we can define various

nonhistorical relations between evolving objects. We can compare species for phenotypic similarity and language varieties for typological similarity. We can compare species for the similarity of their habitat, and language varieties for their geographic closeness. Although these similarities can give us hints regarding deeper historical relations, they are neither a necessary nor a sufficient condition for them.

Evolutionary biology has a rich terminological framework describing fundamental historical relations between genes and morphological characters. Discussions regarding the epistemological and ontological aspects of these relations are frequent and fruitful (Jensen 2001; Koonin 2001; Petsko 2001; Sonnhammer and Koonin 2002; Morrison 2015). In historical linguistics, terminological questions regarding historical relations have occasionally been raised in the past (Katičić 1966; Arapov and Xerc 1974), and recent discussions about the cognacy of grammatical features in historical syntax have emerged (Campbell and Harris 2002; Barðdal and Eythórsson 2012; Walkden 2013). In quantitative applications, however, the fundamental historical relations between words, morphemes, or grammatical features are usually assumed to be self-evident, not deserving specific attention. As a result, our traditional terminology dealing with relatedness, inheritance, and descent is often used imprecisely, frequently leading to confusion in quantitative applications. Computational approaches in historical linguistics are often based on software originally designed for bioinformatics. Scholars justify the use of bioinformatics software in linguistics by drawing analogies between historical relations in the two disciplines. Unfortunately, these analogies often ignore the peculiarities of biological evolution and language history. Instead, they offer a simplified mapping between terms in both disciplines and disregard the underlying processes.

In the following, I will try to illustrate the problems in phylogenetic reconstruction in more detail. I will try to show that the models which are currently used to infer phylogenies from linguistic data suffer from a loss of valid information arising from the superficial analogy between homology and cognacy and a simplification of the processes underlying lexical change. Since terminological misunderstandings are the core of the problem, I will first carry out a brief comparison of biological and linguistic terminology on historical relations, pointing to similarities and differences in the two fields (Section 2). By discussing the complexities of lexical change, I will point to further pitfalls that should be avoided when modeling lexical change with biological software (Section 3). I will then propose improvements to the models currently used in computational historical linguistics (Section 4), and illustrate for a small lexical dataset of Chinese dialects how complex historical relations between words can be modeled in computational approaches to phylogenetic reconstruction (Section 5).

## 2. Terminology for historical relations in biology and linguistics

Scholars have often compared biological and linguistic terminology (Gray 2005; Croft 2008; Pagel 2009; Geisler and List 2013). The analogies that have been made are, however, not necessarily very precise. This becomes especially evident in the analogies drawn between the terms which are used to describe historical relations between evolving objects in both fields. The most popular analogy in this context is that between *homology* in biology and *cognacy* in linguistics (Pagel 2009). In the following, I will carry out a detailed comparison between the terminology used in both fields, thereby showing that the analogy between homology and cognacy is essentially misleading.

### 2.1 Homology

Homology is a fundamental concept in evolutionary biology, designating a 'relationship of common descent between any entities, without further specification of the evolutionary scenario' (Koonin 2005: 311). The term was first defined by Richard Owen (1804–92), who distinguished 'homologues', as 'the same organ in different animals under every variety of form and function' (Owen 1843: 379), from 'analogues' as an 'organ in one animal which has the same function as another part or organ in a different animal' (Owen 1843: 374). Homology is a very general historical relation between evolving objects. It does not specify the process from which the relation originated. Geneticists distinguish three subtypes of homology based on processes underlying the homology of genes in molecular evolution: *orthology*, *paralogy*, and *xenology*. Orthology refers to 'genes related via speciation' (Koonin 2005: 311), paralogy refers to 'genes related via duplication' (Koonin 2005: 311), and xenology refers to genes 'whose history, since their common ancestor, involves an interspecies (horizontal) transfer of the genetic material for at least one of those characters' (Fitch 2000: 229).

In a paper from 1970, Fitch suggested to distinguish two kinds of homology in molecular evolution: homology as the 'result of speciation so that the history of the gene reflects the history of the species' should be called 'orthology', and homology as the 'result of gene duplication so that both copies have descended side by

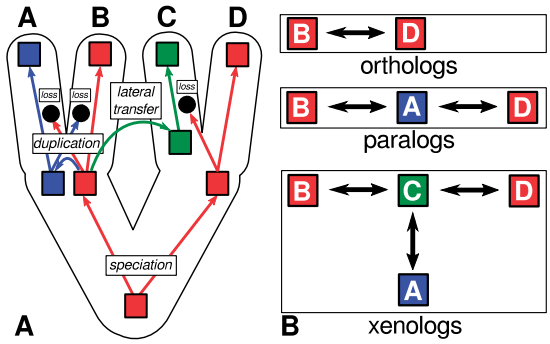


**Figure 1.** Subtypes of homology in molecular biology. Three processes, *speciation*, *duplication*, and *lateral transfer* underly the three basic types of homology in molecular evolution. The processes are illustrated in (A), the resulting relations are illustrated in (B).

side during the history of an organism' should be called 'paralogy' (Fitch 1970: 113). First evidence that genome evolution does not only involve the mutation of individual genes but also the duplication of genes as a whole was reported in the 1930s (Zhang 2003; Taylor and Raes 2004).

In 1983, Gray suggested to use the term *xenology* as a third subtype of homology in order to distinguish those cases in which genes are homologous, but neither orthologous nor paralogous, since 'cells and organisms have acquired foreign genes in the past' (Gray and Fitch 1983: 64). It is now a well-established fact that prokaryotes (bacteria) may acquire genetic material from 'their neighborhood or [...] environment and incorporate it into their genomes' (Nelson-Sathi et al. 2013: 166). Lateral gene transfer processes were first detected and described in the 1950s (Freeman 1951). Only 30 years later, however, scholars began to emphasize the importance of lateral gene transfer for microbial evolution (Syvanen 1985). Figure 1 contrasts the three basic processes of speciation, duplication, and lateral transfer with the resulting historical relations in evolutionary biology.

## 2.2 Cognacy

In historical linguistics, the only relation which is explicitly defined is *cognacy* (also called *cognation*). Cognacy usually refers to words related via 'descent from a common ancestor' (Trask 2000: 63) and it is strictly distinguished from descent involving lateral transfer (*borrowing*). The term cognacy itself, however, covers both direct and indirect descent. Hence, German *Zahn* 'tooth' is cognate with English *tooth*, as is German *Kopf* 'head' with English *cup*, and German *Getränk* 'drink' with English *drink*, although the historical processes that shaped the present appearance of these three word pairs are quite different: apart from the sound shape, *Zahn* and *tooth* have regularly developed from Proto-Germanic **tan Þ* (Kroonen 2013: 509f); *Kopf* and *cup* both go back to Proto-Germanic **kuppa-* 'vessel' (Pfeifer 1993; Kluge and Seebold 2002),[1] but the meaning of the German word has changed greatly; *Getränk* and *drink* go ultimately back to Proto-Germanic **drinkan* 'to drink' (Kroonen 2013: 100f), but the German noun was built as a collective (with prefix *Ge-*) from the nominalized form of the verb (Pfeifer 1993), while the English noun was directly built from the verb. The nominalized form, Proto-Germanic **dranka-* is still reflected in German *Trank* 'potion'. Thus, of the three examples of cognate words, only the first would qualify as having evolved by direct inheritance. Starostin (2013: 140) suggests to distinguish 'etymological cognacy' from 'lexicostatistical cognacy', the former denoting words whose 'forms go back to the same protoform', and the latter denoting words whose 'meanings go back to the same meaning in the proto-language as well'. Trask (2000: 234) suggests the term *oblique cognacy* to label cases in which 'two or more words in related languages [...] continue alternant forms of a single root in the ancestral language', but this term is rarely used and most of the time linguists simply use the term *cognacy* without further specifying what they actually mean.

## 2.3 Beyond homology and cognacy

In an earlier paper (List 2014: 38–46) I abstracted from the processes underlying the historical relations between genes to contrast the biological and the linguistic terminology. In this comparison, I took *common descent* as the most basic relation, with homology as a direct counterpart. The term 'common descent' may be a bit misleading, but what I had in mind by then were all forms of *historical relations*, including those resulting from lateral transfer. Common descent was further subdivided into *direct common descent* (corresponding to orthology), *indirect common descent* (corresponding to paralogy), and *common descent involving lateral transfer* (corresponding to xenology). I then contrasted the abstract relations and the biological terminology with the terminology currently found in linguistics, thereby pointing to missing slots in the linguistic terminology, for which new terms are proposed. Table 1 illustrates this comparison by contrasting the abstract basic

1 Most likely the word is an early borrowing from Latin which happened before the split of English and German (see Pfeifer 1993).

**Table 1.** Comparing biological and linguistic terminology for historical relatedness (with modifications taken from List 2014). Terms in red are suggested to make up for missing terminology in historical linguistics

| Historical relations | | Terminology | | | | |
|---|---|---|---|---|---|---|
| | | Biology | | Linguistics | | |
| Common descent | Direct | Homology | Orthology | Etymological relation | Cognacy | Direct cognacy |
| | Indirect | | Paralogy | | | Indirect cognacy |
| | Involving lateral transfer | | Xenology | | Indirect etymological relation | |

relations with the terminology in biology and linguistics. Relations for which proper terms are missing in linguistics and for which I proposed new terms are colored in red (List 2014: 44). As one can easily see from the table, historical linguistics does not offer direct counterparts for the abstract relations underlying *homology*, *orthology*, and *xenology* in evolutionary biology. Cognacy in historical linguistics is often deemed to be identical with homology in evolutionary biology (Gray 2005; Pagel 2009), but if we follow the comparison, this is only true if one ignores common descent involving lateral transfer, since borrowings are explicitly excluded from the classical definition of cognacy in historical linguistics (Trask 2000: 63).

As we can see from the table, linguistics lacks a proper term for a historical relation between words regardless of whether they are inherited or borrowed (homology in biology, etymological relation according to Table 1). There is also no term denoting the relation between words of which one has been borrowed during its history (xenology in biology). This does not mean, of course, that the relations do not occur in the linguistic domain. Lateral transfer, the process underlying the relation of xenology in molecular biology is also common in language history.[2] In contrast to a relation between two words which involves lateral transfer, the term *borrowing* refers to distinct processes involving a donor and a recipient. As an example for such a relation, consider the words German *kurz* 'short' and English *short* (List 2014: 40). These words are not cognate. German *kurz* is a borrowing from Latin *curtus* 'mutilated' (Pfeifer 1993), but English *short* probably goes back to Proto-Indo-European **(s)sker-* 'cut off' (Rix et al. 2001), and so does Latin *curtus* (Vaan, 2008). The specific history behind these relations is illustrated in Fig. 2. Since German *kurz* was borrowed early from Latin, we cannot say that *kurz* has been borrowed from French *court*, but we also cannot say that both words are cognate. Yet since both words share a common history, it would be likewise wrong to label them as unrelated, in lack of a proper terminology.

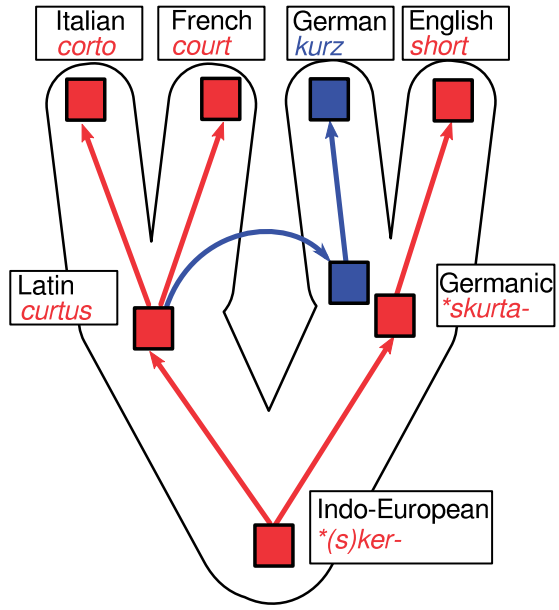


**Figure 2.** Complex historical relations between reflexes of Proto-Indo-European **(s)ker-* 'cut off'.

## 3. Modeling lexical change

In the previous section, I have introduced the basic terminology which biologists and linguists use to denote specific relations between evolving objects. I have then presented an earlier approach of mine (List 2014), where I used the distinctions made in the biological domain in order to introduce new terms for specific historical relations between words. On the first look, the approach seems justified, and the proposed analogies between biological and linguistic relations seem to be fruitful. When looking into the details, however, it becomes clear that important questions are left unanswered. While it is obvious that cognacy in linguistics is not the same as homology in biology, it is less clear how we should understand the idea of *direct* and *indirect* cognacy.

2 We should, of course, be careful with analogies, and it is clear that the specific processes of lexical borrowing are completely different from the processes of lateral gene transfer in biology. On an abstract level, however, the analogy between lateral gene transfer and lexical borrowing holds, in so far as both processes involve the direct transfer of material between evolving objects.

What exactly is meant to be indirect here? Is it the fact that words differ in meaning, thus being akin to words which are *root-cognate* but not *lexicostatistically cognate*, following the distinction of Starostin (2013: 140), or should we instead concentrate on morphological differences, thus following the notion of *oblique cognacy* proposed by Trask (2000: 234)? And how does the idea of 'indirect descent' relate to paralogy and the process of gene duplication in biology? In the following, I will try to show that we need to go beyond my earlier proposal in order to develop a satisfying model of lexical change that can be used for phylogenetic reconstruction.

### 3.1 Degrees of cognacy

Morrison (2015: 50) points to the relative character of homology in evolutionary biology in emphasizing that evolving objects can exhibit homology at different levels, which may even be independent of each other:

> *The classic example is the comparison of bird wings and bat wings. These are homologous as forelimbs (structures), which are general throughout the tetrapods, but they are not homologous as wings (functions), because they represent independent modifications of those forelimbs in the ancestors of birds and bats. (Morrison 2015: 50)*

We can find similar situations in linguistics: if we consider words for 'to give' in the four Romance languages Portuguese, Spanish, Provencal, and French, we can state that both Portuguese *dar* and Spanish *dar* are homologous, as are Provencal *douna* and French *donner*. The former go back to the Latin word *dare* 'to give', the latter go back to the Latin word *dōnāre* 'to gift (give as a present)'. In times when Latin was spoken, both *dare* and *dōnāre* were clearly separated words denoting clearly separated concepts and being used in clearly separated contexts. The verb *dōnāre* itself was derived from Latin *dōnum* 'present, gift'. Similar to English where nouns can be easily used as verbs, Latin allowed for specific morphological processes to turn nouns into verbs. What the ancient Romans were not aware of is that Latin *dōnum* 'gift' and Latin *dare* 'to give' themselve go back to a common word form. This was no longer evident in Latin, but it was in Proto-Indo-European, the ancestor of the Latin language. Thus, Latin *dare* goes back to Proto-Indo-European *$deh_3$- 'to give', and Latin *dōnum* goes back to Proto-Indo-European * $deh_3$ - *no*- 'that what is given (the gift)' (Meiser 1998). The word form *$deh_3$ - *no*- is a regular derivation from *$deh_3$-, so on the Indo-European level, both forms are *homologous*, since one is derived from the other. This means in turn, that Latin *dare* and *dōnum* are also homologues, since

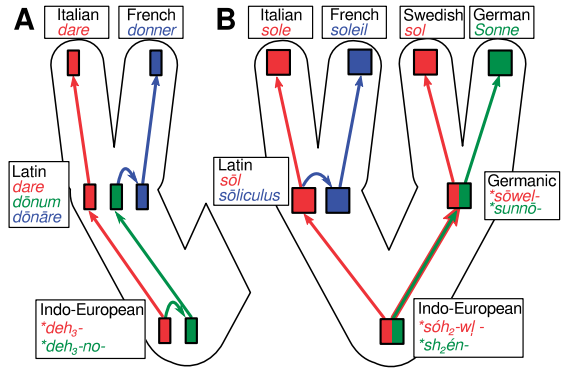


**Figure 3.** Degrees of cognacy in Indo-European language history: the development of words meaning 'to give' from Proto-Indo-European via Latin to Italian and French (A), and the development of words meaning 'sun' in from Proto-Indo-European to Italian, French, Swedish, and German (B).

they are the residual forms of the two homologous words in Proto-Indo-European. And since Latin *dōnāre* is a regular derivation of *dōnum*, it means, again, that Latin *dare* and *dōnāre* are also homologous, as are the words in the four descendant languages, Portuguese *dar*, Spanish *dar*, Provencal *douna*, and French *donner*. Depending on the time depth we apply, we will arrive at different homology decisions. The history of the words is depicted in Fig. 3A.

An even more complex example are words like Italian *sole*, French *soleil*, Swedish *sol*, and German *Sonne*, all meaning 'sun'. Indo-European scholars assume that the Proto-Indo-European word for sun had a complex, stem-alternating paradigm with two different base forms, one for nominative and accusative case *$séh_2u̯el$-, and one for the oblique cases, *$sh_2én$- (Wodtko et al. 2008: 606). Proto-Germanic inherited this paradigm completely (*sōel- versus *sunnōn, Kroonen 2013: 463f), but it was simplified via the process known as *analogy* in historical linguistics, and the nominative stem was taken as the base form in Latin *sōl* (Meyer-Lüebke 1911: §8059). In Swedish and German, the complex base form was also simplified, but in different directions, with the Swedish form taking the nominative stem as the basis of analogy, and the German form taking the oblique stem. While Italian *sole* is the regular reflex of Latin *sōl*, French *soleil* goes directly back to Latin *sōliculus* 'small sun', a Latin diminutive of *sol* (Meyer-Lüebke 1911: §8067). From this perspective, Italian *sole* is more closely related to Swedish *sol* than to French *soleil*, although French and Italian are, of course, much closer genetically related than are Swedish and Italian. The history of the reflexes of the Indo-European word for 'sun' is depicted in Fig. 3B.

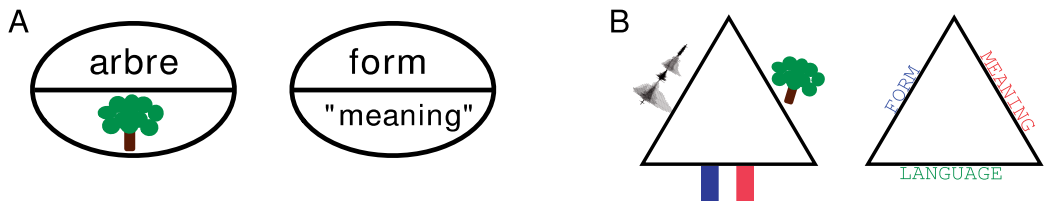


**Figure 4.** The different dimensions of the linguistic sign: (A) Shows the classical model after Saussure (1916). (B) Shows an extended sign model in which the language, the system in which a sign is used was added as a third component.

### 3.2 Dimensions of lexical change

In a very simple language model, the lexicon of a language can be seen as a bag of words. A word is further defined by two aspects: its *form* and its *meaning*. Thus, the French word *arbre* can be defined by its written form *arbre* or its phonetic form [ɑʀbʀə], and its meaning 'tree'. This is reflected in the famous sign model of Ferdinand de Saussure (1857–1913, Saussure 1916), which I have reproduced in Fig. 4A. In order to emphasize the importance of the two aspects, linguists often say that form and meaning of a word are like two sides of the same coin, but we should not forget that a word is only a word if it belongs to a certain language. From the perspective of the German or the English language, for example, the sound chain [ɑʀbʀə] is just meaningless. So instead of two major aspects of a word, we may better talk of three major aspects: *form*, *meaning*, and *language* (Ternes 1987: 22f; List 2014: 15–18). As a result, our bilateral sign model becomes a trilateral one, as illustrated in Fig. 4B.

Gévaudan (2007) distinguishes three dimensions of lexical change: The *morphological dimension*, the *semantic dimension*, and the *stratic dimension*. The morphological dimension points to changes in the form of words which are not due to regular sound change. As an example, consider German *Getränk* 'drink' and its ancestral form Old High German *tranc* 'drink'. While the meaning of the word is the same, the German word *Getränk* is a collective derivation of the Old High German source form (Pfeifer 1993). The derivation process involved prefix *Ge-*, and the modification of the main vowel. The semantic dimension is illustrated by changes like the one from Proto-Germanic **kuppa-* 'vessel' to German 'Kopf'. The stratic dimension refers to changes which involve lexical material *outside* the *historical continuum* of a given language (Gévaudan 2007: 141f). In the terminology of Gévaudan (2007: 141f), *stratum* refers to languages as historical continua, and should not be confused with the way the term is used in sociolinguistics, where it refers to language varieties used in certain layers of a linguistic society (Coseriu 1973; Oesterreicher 2001), but rather in opposition to the term *adstratum* in historical and areal linguistics (Gévaudan 2007: 141). Usually, changes along the stratic dimension belong to the class of *borrowing processes*. (Gévaudan 2007: 141–63) argues, however, that processes like onomatopoeia, antonomasy, and folk etymology can also be characterized as processes which involve the stratic dimension of lexical change, since they are based on material which does not stem from the historical continuum of a given language. An example for a simple type of stratic change is English *mountain* which was borrowed from Old French *montaigne* 'mountain'. An example for a more complex type of stratic change is German *Maus* 'mouse (for a computer)' which was not directly transferred from English but rather received a broadened semantic function under the influence of the English word (compare Weinreich (1974: 47–62) and Gévaudan (2007: 143–51) for more details on different types of lexical interference).

Note that these three dimensions of lexical change correspond directly to the three major aspects constituting the linguistic sign: the morphological dimension changes the *form* of a word, the semantic dimension its *meaning*, and the stratic dimension its *language*. Thus, the three dimensions of lexical change, as proposed by Gevaudan find their direct reflection in the major dimensions along which words can vary.

### 3.3 27 Shades of cognacy

When looking at the different historical relations from the perspective of the three dimensions of lexical change, it becomes clear that the new terms I proposed earlier (List 2014) do not necessarily solve our problem of reflecting the different aspects of lexical change and lexical variation adequately. Although it seems justified to point to the difference between cognacy in linguistics and homology in biology, it proposes a problematic analogy between paralogy and indirect cognacy without further specifying how indirect cognacy should be defined in the end. When investigating the different uses of the

**Table 2.** 27 shades of cognacy: the table shows exemplarily how cognacy can be modeled according to the three dimensions of lexical change, highlighting potential analogies in biology.

| | | continuity | | |
|---|---|---|---|---|
| **Relation** | **Biol. Term** | Stratic | Morphological | Semantic |
| traditional notion of cognacy | - | + | +/- | +/- |
| cognacy à la Swadesh (1952, 1955) | - | + | +/- | + |
| direct cognate relation (List 2014) | orthology | + | + | + |
| oblique cognate relation (Trask 2000) | | + | - | +/- |
| etymological relation (List 2014) | homology | +/- | +/- | +/- |
| oblique etymological relation (List 2014) | xenology | - | +/- | +/- |
| ... | ... | ... | ... | ... |

term 'cognacy', for example, it becomes obvious that the differences result from *controlling* for one or more of the three dimensions of lexical change proposed by Gévaudan (2007).[3] The notion of cognacy of a classical Indo-Europeanist, for example, controls the stratic dimension by requiring *stratic continuity* (no borrowing), but at the same time it is indifferent regarding the other two dimensions. This is what Starostin (2013: 140) called 'etymological cognacy'. Cognacy à la Swadesh (especially Swadesh 1952,1955), as we know it from lexicostatistics (Swadesh 1952, 1955) and its modern derivations (Gray and Atkinson 2003), is indifferent regarding *morphological continuity*, but controls the semantic and the stratic dimensions by only considering words that have the same meaning and have not been borrowed. This is what Starostin, (2013: 140) called 'lexicostatistical cognacy'.

'Traditional cognacy' and 'cognacy à la Swadesh', however, are but two ways to control for the three dimensions of lexical variation, and one can easily think of more perspectives on historical relations between words, including the terminology that is used in evolutionary biology. In Table 2, I have attempted to illustrate in which way the different terms, including the biological terms of homology, orthology, and xenology, cover processes by controlling each for one or more of the three dimensions of lexical change (with + indicating that continuity is required, − indicating that change is required, and +/− indicating indifference). Note that paralogy was not included in the comparison, since the process of gene duplication is a very specific event that probably has no fruitful analogy in historical linguistics. Contrasting the different dimensions of lexical change with the terminology used to refer to different relations between words shows the arbitrariness of the traditional linguistic terminology. Why do we only cover two out of $3 * 3 * 3 = 27$ different possible types? Why do we only control by requiring continuity, not change? It also shows the fundamental difference between change processes in linguistics and biology.

## 4. Models of lexical change in phylogenetic reconstruction

In the previous sections, I have tried to show that not only the terminology that we use to denote historical relations between evolving entities in linguistics and biology shows some important differences, but also that the processes underlying lexical change in language history are very particular, involving three major dimensions of lexical variation which themselves can be further subdivided into a multitude of minor process types.[4] In the following, I will try to illustrate how our models can be modified in order to account for more complex historical relations between words.

### 4.1 Gain loss models and morphological variation

The majority of automatic methods for phylogenetic reconstruction in historical linguistics employ lexical data to infer language phylogenies. When employing these

3 Note that, in this context, 'controlling' for a dimension means to consider only those historically related words in which *no variation* along that very dimension occurred *during their history since separation*. If we compare French *soleil* 'sun' with Italian *sole* 'sun', for example, we would need to state that the French word changed its meaning from *small sun* to *sun*, and although both forms are identical regarding their synchronic meaning, their history involves variation along the semantic dimension (see Starostin 2013 for more examples on cases of *unilateral independent semantic development*). In practice, when linguists prepare lexicostatistical databases, however, controlling for meaning is usually reduced to checking for identity along a given dimension. It is clear that this can be problematic. In the absence of counterevidence the majority of linguists would probably assume that meaning identity in cognate word forms is good evidence that no semantic change happened since the separation of the forms, but it is obvious that semantic identity is only a necessary for semantic continuity since separation.

4 Already a brief overview of some classical work on the complexities of semantic change (Wilkins 1996), morphological change (Koch 1996), and stratic change (Weinreich 1974) shows that the three-dimensional model of lexical change only touches the tip of the huge iceberg of lexical change.

**Table 3.** Lexicostatistical scheme of data-encoding and the creation of presence-absence matrices. The table shows how lexicostatistical word lists are produced, how cognates are assigned to words by using numerical identifiers, and how the data are then converted into binary presence absence matrices for the purpose of phylogenetic comparison. Note that the proto-form which is given for each cognate set in the table below is not necessarily included in lexicostatistical datasets, but it, nevertheless, is implicitly assumed.

| Basic Concept | German | ID | English | ID | Italian | ID | French | ID |
|---|---|---|---|---|---|---|---|---|
| HAND | Hand | 1 | hand | 1 | mano | 2 | main | 2 |
| BLOOD | Blut | 3 | blood | 3 | sangue | 4 | sang | 4 |
| HEAD | Kopf | 5 | head | 6 | testa | 7 | tête | 7 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

| ID | Proto-Form | Basic Concept | German | English | Italian | French |
|---|---|---|---|---|---|---|
| 1 | PGM **xanda-* | HAND | 1 | 1 | 0 | 0 |
| 2 | LAT *mānus* | HAND | 0 | 0 | 1 | 1 |
| 3 | PGM **bloda-* | BLOOD | 1 | 1 | 0 | 0 |
| 4 | LAT *sanguis* | BLOOD | 0 | 0 | 1 | 1 |
| 5 | PGM **kuppa-* | HEAD | 1 | 0 | 0 | 0 |
| 6 | PGM **xawbda-* | HEAD | 0 | 1 | 0 | 0 |
| 7 | LAT *tēsta* | HEAD | 0 | 0 | 1 | 1 |
| ... | ... | ... | ... | ... | ... | ... |

methods, it is important to specify a model of lexical change that the algorithms can use to infer the trees or the networks that fit the data best. Most datasets employ a lexicostatistical scheme of data-coding (Dyen et al. 1992; Ringe et al. 2002; Greenhill et al. 2008; Bouckaert et al. 2012; Greenhill 2015). This means, that they are based on concept lists of 100 and more items which are translated into the languages under investigation. By comparing all translations in each concept slot with each other, linguists then annotate which words are cognate. The notion of cognacy that is underlying these databases is usually the notion of 'cognacy à la Swadesh' in Table 2, that is, annotators try to filter out borrowings, consider only semantically identical items, and do not necessarily regard morphological variation.

The methods which are then used to analyze the data, be they based on probabilistic approaches (Felsenstein 1981; Huelsenbeck et al. 2001), or parsimony (Fitch 1971; Sankoff 1975), are almost exclusively based on gain–loss models of lexical change (Pagel 2009). They reduce the change of phylogenetic characters to processes of gain and loss and essentially assume that during evolution a language can either gain a new word or loose an existing one. In these models, each phylogenetic character has only two states, presence, or absence, and presence–absence matrices of cognate sets are fed to the algorithms in order to infer language phylogenies. Presence–absence matrices are retrieved from the original data by breaking up the semantic slots into sets of cognate words, and listing for each language whether it has a word belonging to the respective cognate set or not (Atkinson and Gray 2006). This way of data preparation and encoding is further illustrated in Table 3.

The binary coding practice has strong consequences, since it is vulnerable to historical word relations with variation along the semantic and the morphological dimension. First, the general procedure by which lexicostatistical data is binary encoded and concepts are split into several independent characters creates dependencies which cannot be observed by the algorithms. It deprives the analysis of the essential criterion for gain and loss, since presence and absence are defined with respect to meaning identity. Gain and loss need to be essentially interpreted as gain and loss with respect to a certain concept slot, not with respect to the entire language. The loss of a word means that the word is no longer used to express a certain meaning, and the gain of a word implies that a new word is used to express a certain meaning. Yet since meaning is discarded by the binarisation procedure (see Table 3), the models are given no clue to handle instances of parallel semantic shift. A more realistic gain–loss analysis should include a larger sample of words and annotate cognates regardless of differences in meaning (Michael et al. 2015).

Second, the lexicostatistical coding practice is vulnerable with respect to morphological change, since morphological variation is deliberately ignored when assigning words to cognate sets. This was not the case in the early days of lexicostatistics. Hattori (1961), for example, distinguished clearly between true 'orthologues' and morphologically derived words. Recalling the example of Italian *dare* and French *donner* given in Fig. 3, it is clear that we can annotate the words quite differently, depending not only on the "shade" of cognacy we choose, but also on the desired depth of analysis. In current practice, words like *dare* and *donner* are usually assigned to the same cognate set, and their morphological differences are ignored.[5] When annotating the words, however, we should ask ourselves which kind of annotation would be the best for the underlying model that we use. From this perspective, we would do best in coding Italian *dare* and French *donner* as being not cognate, since by the time that *donner* replaced earlier *dare* in the ancestor of French, the word *dare* was lost with respect to the meaning 'to give', and the word *donner* was gained.

5 Compare the coding in the Indo-European Lexical Cognacy Database at http://ielex.mpi.nl/cognate/405/, version accessed on 2016-04-08 available at WebCite: http://www.webcitation.org/6dGAxAG9r.

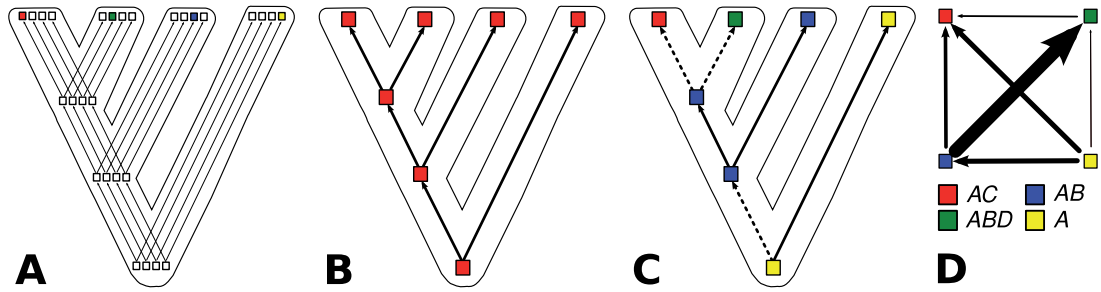


**Figure 5.** From gain–loss models to weighted directed character-state transitions: (A) Shows a strict approach in which four partially related compound words (as show at the bottom of D) are modeled as four different characters. (B) Shows the consequences of a lumping approach when partially cognate words are treated as fully cognate in binary presence–absence models. (C) Shows weighted directed character–state transitions, based on known transition tendencies displayed at the top of (D), with arrows indicating directions and edge width indicating the relative strength of transition tendencies.

**Table 4.** Complex etymological structure in word compounds. The table shows partial etymological relations of words for 'moon' in four Chinese dialects. Dialect data Hóu (2004), Middle Chinese (MC) readings follow Baxter (1992) with modifications.

| Variety | Form | Character | Etymological structure | | | |
|---|---|---|---|---|---|---|
| | | | MC *ŋiot 月 | MC *kwaŋ 光 | MC *bjut 佛 | MC *ljaŋɦ 亮 |
| Fúzhōu 福州 | ŋuoʔ$^{5}$ | 月 | ŋ u o ʔ $^{5}$ | | | |
| Měixiàn 梅縣 | ŋiat$^{5}$ kuoŋ$^{44}$ | 月光 | ŋ i a t $^{5}$ | k u o ŋ $^{44}$ | | |
| Wēnzhōu 溫州 | ɲy$^{21}$ kuɔ$^{35}$ vai$^{13}$ | 月光佛 | ɲ - y - $^{21}$ | k u ɔ - $^{35}$ | v a i $^{13}$ | |
| Běijīng 北京 | yɛ$^{51}$ liaŋ$^{1}$ | 月亮 | − y ɛ - $^{51}$ | | | l i a ŋ $^{1}$ |

The problem of morphological variation in lexicostatistical datasets becomes even more evident when looking at more specific processes of morphological change like *compounding*. While compounding is less characteristic for the Indo-European language familiy (at least as far as the stable parts of the lexicon are concerned), it plays an important role in the Sino-Tibetan language family (Matisoff 2000: 341f; Chung et al. 2014; List 2015: 56–58). In the Chinese dialects, for example, the majority of words is only indirectly related, as illustrated in Table 4 where the words for 'moon' in four Chinese dialects share the same base morpheme, but differ regarding the further parts of their compounds. When investigating these patterns, we can immediately infer processes of lexical change that link these patterns. Fúzhōu [ŋuoʔ$^{5}$] 月, for example, reflects the oldest stage in which Chinese was still predominantly monosyllabic. Měixiàn [ŋiat$^{5}$ kuoŋ$^{44}$] 月光 reflects a younger stage in which bisyllabic structures were gaining ground, and Wēnzhōu [ɲjy$^{21}$kuɔ$^{35}$vai$^{13}$] 月光佛 reflects an even later stage, since it builds on the form in Měixiàn, adding a suffix that marks nominalization (compare Wēnzhōu [ɲji$^{21}$dʏu$^{35}$vai$^{13}$] 日頭佛 'sun').[6] In the 'classical' lexicostatistical view of cognacy and the 'classical' models of word gain and word loss, these processes are all ignored, although they may bear important phylogenetic information. One would either label all four words as cognate, since they share the same base morpheme (Satterthwaite-Phillips 2011: 95–103), or label them all as not being cognate, since their parts to not match completely (Ben Hamed and Wang 2006; Gates 2012: 51). If we want to model the evolution of the four words for 'moon' in the four dialects realistically, neither of the two encoding practices will be of use. In both cases, all phylogenetic signal will be lost and the analysis cannot tell us how the words really developed (see Fig. 5A and B).

## 4.2 From binary to multistate models

In principle, phylogenetic methods can handle semantic and morphological variation sufficiently. All we need to

6 Note that in this case, as in general when dealing with lexical change in a classical lexicostatistical framework, sound change is ignored as a factor of change, since regular sound change involves the sound system and not individual phonetic material (Gévaudan, 2007: 14).

do is to switch from binary gain–loss models to multistate models. In a binary state model each character can only be present or absent in a given language, like the cognate set 1 in Table 3, for example, which is present in German and English but absent in Italian and Spanish. In a multistate model, a character cannot only be present or absent, but it can also *vary* among languages and occur in different shapes. Instead of labeling French *donner* and Italian *dar* either as exclusively cognate or as exlusively noncognate, we could assign both words to the same character but assign them different states. In this way, we could handle both variation along the semantic and the morphological dimension of lexical change. If we can further determine how likely it is for the character to switch from one particular state to another, we can force our algorithms to prefer certain transitions and to ignore others. In the case of the Chinese words for 'moon' in Table 4, for example, we already saw that Měixiàn [ŋiat$^{5}$ kuoŋ$^{44}$] 月光 is particularly close to Wēnzhōu [ŋjy$^{21}$kuɔ$^{35}$vai$^{13}$] 月光佛, since the latter was only extended by one suffix. When comparing the Wēnzhōu form with the form [ŋuoʔ$^{5}$] 月 in Fúzhōu, we can further easily say that the transition from the Fúzhōu form to the Měixiàn form should be easier to accomplish than the direct transition to the Wēnzhōu form. If we further know that the process we are dealing with has strong *unidirectional tendencies*, as it is the case for many processes of sound change and grammaticalization (Haspelmath 2004), but also in inflectional morphology (Wurzel 1985), and potentially even in analogy (Jacques 2016), we can model this by using *irreversible models* in our analyses (Huelsenbeck et al. 2002; Bohl and Lancaster 2003).

In a parsimony framework of phylogenetic reconstruction (Fitch 1971; Sankoff 1975), the difficulty of switching between the different states of a character is handled by defining specific *weights* for character state transitions. If we further know that the process we are dealing with has strong unidirectional tendencies, we can model this by assigning *asymmetric weights* for the transition preferences between the states of a character. The differences between gain–loss models and multistate models allowing for asymmetric transition preferences in a parsimony framework are exemplified in Fig. 5, but multistates and asymmetric transition tendencies can essentially also be handled in probabilistic frameworks.

## 5. Using improved models to study Chinese dialect history

In order to illustrate the benefits of improved models for lexical change, I have prepared a small experiment on Chinese dialect history. In this experiment, I test how well different models of lexical change with varying degrees of complexity perform on the task of *semantic reconstruction*. In classical historical linguistics, semantic reconstruction seeks to infer the original meaning of a set of cognate words (Fox 1995: 115–6). The experiment I designed follows lexicostatistical approaches in which semantic reconstruction seeks to identify the word form which was used to express a certain concept in an ancestral language (Kassian et al. 2015: 304–6). In this context, semantic reconstruction can be treated as a specific type of *ancestral state reconstruction* (Pagel 1999) applied to lexicostatistical data. The starting point is a lexicostatistical wordlist, consisting of a list of concepts which are translated into a set of language varieties. Concepts comprise phylogenetic characters, and the counterparts of the concepts in the respective language varieties reflect different states of the characters. Semantic reconstruction starts from a *reference phylogeny* (a phylogenetic tree) and tries to infer which character state was present at the root. Chinese is attested through its contemporary dialects, whose diversity is at least comparable to that of the Romance languages (Wang 1997), but also in ancient texts predating the diversification of the modern dialect varieties by several hundred years.[7] Therefore, in the majority of cases, there is independent evidence regarding the words which were originally used to express a given concept. For this reason, Chinese is an ideal candidate to test the performance of different models of lexical change.

7 There is some disagreement among Chinese linguists regarding the exact dating of the ancestor of all Chinese dialects. Some scholars assume that the modern dialects developed from a *koine* spoken in the early Táng 唐 dynasty (618–907 AD) around 600 AD (Karlgren 1954; Pulleyblank 1984). Other scholars propose an earlier diversification. Assuming that the very conservative Mǐn 閩 dialect group had much earlier split off from the rest of Chinese (Norman and Coblin 1995; Handel 2010), they place their common ancestor in the late Hàn 漢 dynasty (206 BC–220 AD) some time around 200 AD. Nevertheless, with ancient Chinese texts dating back to 1000 BC and earlier, with rich collections of classical texts being available from the sixth century BC onwards, Ancient Chinese is clearly ancestral to all Chinese dialects, as is also reflected in its sound system (Baxter and Sagart 2014).

**Table 5.** The concepts selected for the study

| | | | | | |
|---|---|---|---|---|---|
| 1. ash / 灰 | 2. back / 背 | 3. belly / 腹 | 4. bird / 鸟 | 5. bone / 骨 | 6. claw / 爪 |
| 7. cloud / 云 | 8. day / 天 | 9. dog / 犬 | 10. ear / 耳 | 11. earth / 地 | 12. eat / 食 |
| 13. egg / 卵 | 14. eye / 目 | 15. fire / 火 | 16. flesh / 肉 | 17. flower / 花 | 18. fog / 雾 |
| 19. fruit / 果 | 20. guts / 肠 | 21. hand / 手 | 22. heart / 心 | 23. horn / 角 | 24. ice / 冰 |
| 25. knee / 膝 | 26. lake / 湖 | 27. leaf / 叶 | 28. leg / 脚 | 29. liver / 肝 | 30. louse / 虱 |
| 31. man / 男 | 32. moon / 月 | 33. mouth / 口 | 34. name / 名 | 35. neck / 颈 | 36. night / 夜 |
| 37. nose / 鼻 | 38. path / 路 | 39. person / 人 | 40. river / 江 | 41. rope / 索, 绳 | 42. sand / 沙 |
| 43. seed / 种 | 44. skin / 皮 | 45. sky / 天 | 46. smoke / 烟 | 47. snake / 蛇 | 48. star / 星 |
| 49. stone / 石 | 50. sun / 日 | 51. tail / 尾 | 52. tongue / 舌 | 53. tooth / 牙 | 54. water / 水 |
| 55. wing / 翼 | 56. woman / 女 | 57. worm / 虫 | | | |

### 5.1 Materials

The data for the experiment were originally compiled for the study of Ben Hamed and Wang (2006). It comprises 200 concepts translated into 23 Chinese dialect varieties. The concept list is largely identical with the list of 200 items proposed by Swadesh (1952).[8] In the data, partial cognate relations are annotated by listing the 'etymological character' for each morpheme of a word (*běnzì* 本字, Branner 2000: 35). This information is regarded as problematic by some Chinese dialectologists (Branner 2000), since it is not necessarily clear how consistently the morphemes in dialect words are identified. Datasets like the one by Ben Hamed and Wang (2006) are, nevertheless, a useful starting point for experiments on morphological processes in lexical evolution, especially since other collections which list information on partial cognacy in such great detail are not available. For most of the cases, however, we can assume that the assignments are correct. In an earlier study (List 2015), I used the data by Ben Hamed and Wang (2006) and converted it into a machine-readable text format, which I used for this experiment. All data were thoroughly checked and refined, since the partial cognate assignments were not the primary target of my earlier study and therefore only inconsistently converted into text format.

Ben Hamed and Wang (2006) also give the ancestral forms for the concepts in Old Chinese. Since Old Chinese is supposed to be the ancestor of all dialect varieties in the sample, the data can be used as a 'gold standard' to test the accuracy of ancestral state reconstruction methods. Since processes of lexical evolution are quite different for nouns and verbs, with compounding and partial cognacy occurring almost exclusively on nouns, only nouns were considered for this study. Of the 85 concepts denoting nouns in the sample, 28 were excluded. Either the reflexes were all different from the Old Chinese forms and it would be impossible to reconstruct them, or the reflexes were all identical with the Old Chinese form, and reconstruction would be no challenge at all. The 57 forms considered for the experiment are listed in Table 5 along with the supposed ancestral forms in Old Chinese.

Ancestral state reconstruction requires a reference phylogeny as input. Here I build on an earlier approach (List 2015) where I compared reference phylogenies for three independent hypotheses on Chinese dialect history, namely Laurent Sagart's *Arbre des Dialectes Chinois* (Sagart 2011), the *Hànyǔ Fāngyán Shùxíngtú* 漢語方言樹形圖 ('Tree chart of Chinese dialects') by Yóu Rǔjié 游汝傑 (Yóu 1992: 91–106), and Jerry Norman's *Southern Chinese Hypothesis* (Norman 1988: 210–4). These reference phylogenies differ regarding the subgrouping of the seven major dialect groups of Chinese and are based on competing criteria for subgrouping (see List 2015: 36f for details).

### 5.2 Methods

The experiment employs a *parsimony framework* for character transitions (Nunn 2011: 59–63). Parsimony was used for reasons of simplicity and data sparseness. Parsimony applications can be easily implemented from scratch, while there are no available ready-to-use implementations of probabilistic approaches which handle asymmetric transitions between multiple character states. Given the sparseness of the data available for testing, it is also not clear whether probabilistic applications would converge. Four different models of varying complexity were defined for the experiment:

a. BINARY: Character states which are not completely identical in their compound structure are split into sets of binary characters following the classical procedure described in Atkinson and Gray (2006).

8 The list is included into the Concepticon resource (http://concepticon.clld.org, see List et al. 2016).

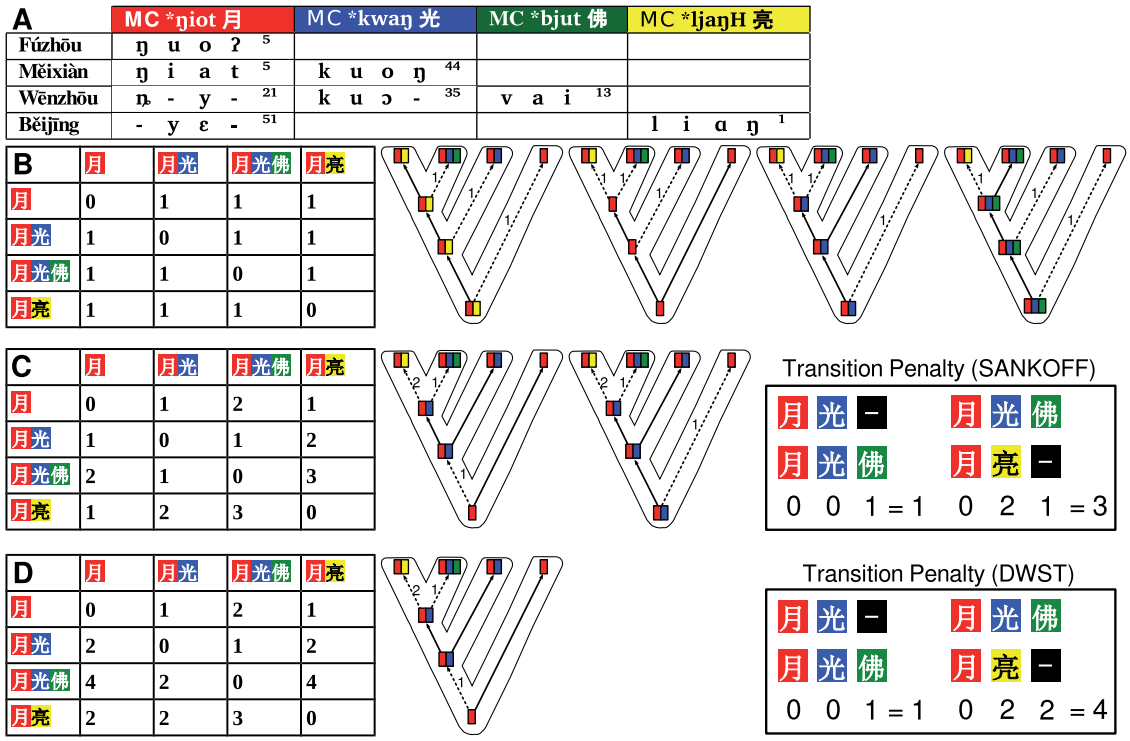


**Figure 6.** Comparing multistate models for lexical change. The figure shows how the evolution of the four words for 'moon' is inferred within a parsimony framework. On top, the etymological structure of the words is displayed, and unique colors are assigned to refer to the morpheme structure in the remainder of the figure (A). On the left, the penalties for character transitions (step matrices) are shown for the FITCH (B), the SANKOFF (C), and the DWST model (D). For SANKOFF and DWST, example calculations for transition penalties are displayed on the right (see also the main text). For each model, all trees with optimal weight are displayed. Dashed edges in the trees indicate a transition involving a change. Numbers on dashed lines denote the weight, as derived from the corresponding matrix of transition penalties on the left.

Character transitions are modeled as a gain–loss process.

b. FITCH (multistate): Lexical evolution is modeled as a process of character transitions with equal weights, following the classical model by Fitch (1971).

c. SANKOFF (multistate, weighted): Lexical evolution is modeled as a process of character transitions with unequal weights, following the classical model by Sankoff (1975).

d. DWST ('directed weighted state-transitions', multistate, weighted, directed): Lexical evolution is modeled as a process of character transitions with unequal weights and in dependence of the direction of the transition.

The BINARY and the FITCH model are straightforward in their implementation. The BINARY model only handles gains and losses with losses being favored over gains. The parsimony weight for gain events was set to 2, and the penalty for loss events was set to 1, since these penalties yielded the most plausible scenarios in earlier experiments on the data (List 2015). The FITCH model gives equal weights to transitions between all states. In the case of SANKOFF and DWST, transitions are weighted differently depending on the character states. Since we lack exhaustive linguistic accounts on processes of compounding in the Chinese dialects, a very simple approach for the computation of the weights was employed. In a first step, the morpheme representation of two words, which is given in Chinese character readings, with identical characters representing cognate morphemes, was aligned using the Needleman–Wunsch algorithm (Needleman and Wunsch 1970). In a second step, it was counted in how many positions the aligned sequences differ. This distance, commonly known as the Hamming distance (Hamming 1950), was further refined by counting substitutions (those instances where two different morphemes are aligned) twice, and insertions and deletions (those instances where a morpheme was aligned with a gap symbol or vice versa) only once.

**Table 6.** Comparing the results for the four analyses and the three reference trees. The first number in the *hits* and the *fails* column indicates the proportion, the second number indicates the absolute values. As mentioned in the text, hits and fails are computed by comparing for all proposed forms reconstructed back to the root whether they are identical with the forms in the gold standard. If they are, this counts as a *hit*, if not, this counts as a *fail*. If more than one form are proposed for a given concept, results are averaged.

| Model | *Arbre* | | *Shùxíngtú* | | *Southern Chinese* | |
|---|---|---|---|---|---|---|
| | Hits | Fails | Hits | Fails | Hits | Fails |
| BINARY | 0.55 / 31.04 | 0.45 / 24.96 | 0.52 / 29.04 | 0.48 / 26.96 | 0.52 / 28.95 | 0.48 / 27.05 |
| FITCH | 0.63 / 35.51 | 0.37 / 20.49 | 0.51 / 28.31 | 0.49 / 27.69 | 0.47 / 26.40 | 0.53 / 29.60 |
| SANKOFF | 0.76 / 42.83 | 0.24 / 13.17 | 0.67 / 37.50 | 0.33 / 18.50 | 0.62 / 34.50 | 0.38 / 21.50 |
| DWST | 0.82 / 45.70 | 0.18 / 10.30 | 0.82 / 46.00 | 0.18 / 10.00 | 0.79 / 44.50 | 0.21 / 11.50 |

In contrast to the SANKOFF model, the computation of weights for the DWST model only reduces the weights for insertions (a gap aligned with a morpheme), but not for deletions. This transition schema accounts for the tendency of *disyllabification* in the history of Chinese, during which most of the monosyllabic words in the Chinese dialects were replaced by bisyllabic compounds. Figure 6 gives examples for the differences in the transition penalties of the multistate-models (FITCH, SANKOFF, and DWST) and the calculation of the transition penalties for the SANKOFF and the DWST model. It is beyond doubt that the models could be further refined, and potentially also trained. For the purpose of the experiment, however, it is advisable to keep the models as abstract as possible. This guarantees that we do not overly fit the models to the data, and it also makes it easier to determine the major factors that determine differences in their performances.

The models and the code to optimize the parsimony score were implemented in Python. The code requires the LingPy software package for quantitative tasks in historical linguistics (List and Moran 2013) to calculate the alignments between the characters states and the transition probabilites. The source code along with the data, the results, and further instructions on how to replicate all analyses presented in this article are provided as supplementary data.

### 5.3 Results

With four different models and three different reference phylogenies, 12 different tests needed to be carried out. In order to evaluate the quality of semantic reconstruction, a simple approach was used. In this approach, one counts the amount of *hits* and *fails*. For each concept, all ancestral forms proposed by a given test were considered and compared with the known forms in the 'gold standard'. If only one form was proposed, this form can either be a hit or a fail, that is, it can either be identical with the form in the gold standard, or not. If multiple forms are proposed by an algorithm, the score is divided among hits and fails, following the proportion of correctly and incorrectly proposed ancestral forms. If, for example, two forms are proposed of which only one is correct, this would be scored as a 50% hit and a 50% fail. The results were evaluated separately for each meaning slot and then averaged across all 57 concepts in the sample.

Table 6 shows the detailed results for all 12 different analyses, including the overall parsimony scores obtained. The DWST model performs best in all respects, regardless of the reference phylogeny. The SANKOFF model outperforms the remaining two models, but only when applied to the *Arbre* reference phylogeny, it comes close to the high scores of the DWST model. Whether the BINARY or the FITCH model performs better is hard to say, given that the differences are minimal on average, and both models seem to rely heavily on the reference phylogeny. What is remarkable is that the DWST model does not only show the highest scores, but also a high resistency regarding the underlying reference phylogeny. According to the analysis by List (2015), the *Arbre* gives a more realistic picture of Chinese dialect history. This is reflected by the highly improved scores of all models (except from DWST) for the *Arbre* phylogeny as opposed to *Shùxíngtú* and *Southern Chinese*.

Parsimony approaches may yield multiple solutions which are all optimal with respect to the transition penalties defined in a model. Depending on the character and tree topology, the amount of optimal scenarios may vary greatly. In the FITCH analyses, for example, the number of possible scenarios for all characters ranges from 1 (for 'ash') to 4 797 (for 'night'). As expected, the

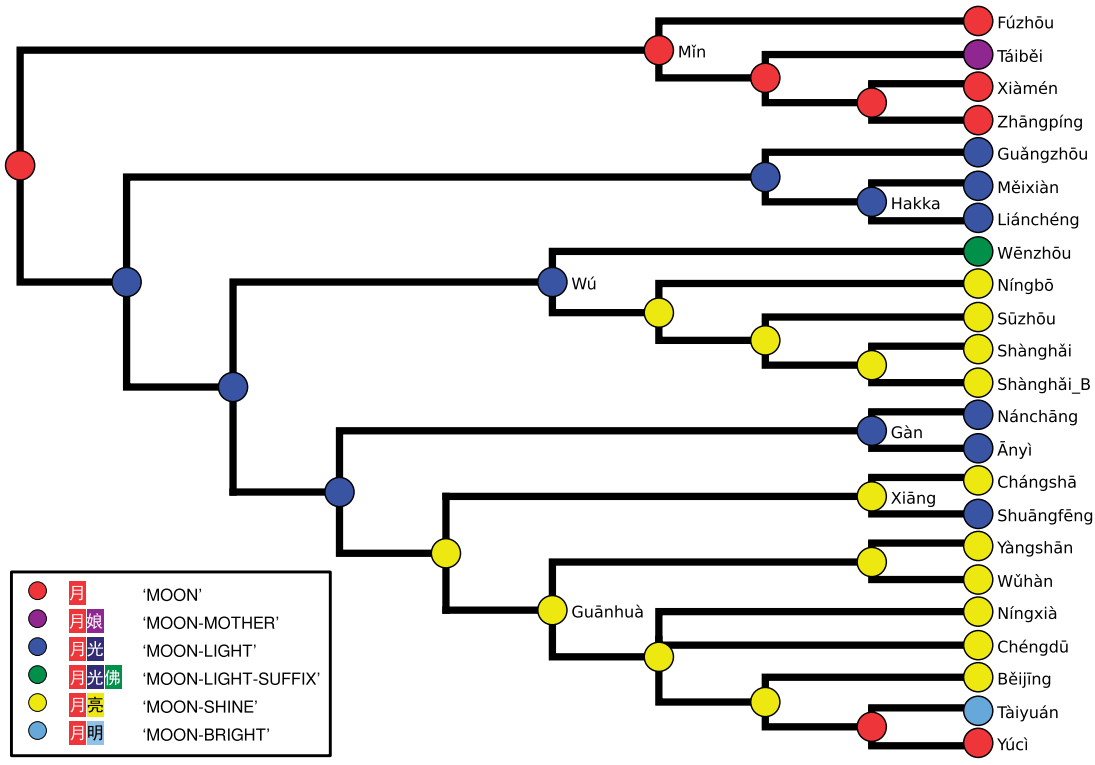


**Figure 7.** One of four optimal scenarios for the development of words for 'moon' along the *Arbre* reference phylogeny.

**Table 7.** Comparing the proposed proto-forms and the number of optimal scenarios based on the *Arbre* reference phylogeny for three exemplary concepts. Forms with an asterisk represent 'hits', that is, forms which are identical with the gold standard.

| Models | 'belly' | | 'ear' | | 'moon' | |
|---|---|---|---|---|---|---|
| | Forms | Scen. | Forms | Scen. | Forms | Scen. |
| FITCH | 肚, 腹老, 腹肚, 腹* | 39 | 耳*, 耳朵, 耳仔, 耳菇, 耳公 | 34 | 月*, 月光 | 48 |
| SANKOFF | 肚*, 腹肚, 腹 | 5 | 耳* | 3 | 月*, 月光 | 8 |
| DWST | 腹*, 肚 | 2 | 耳* | 1 | 月* | 4 |

number of possible scenarios decreases when increasing the complexity of the models. This is shown in Table 7 where the proposed proto-forms and the number of possible scenarios for the analysis of three concepts using the three multistate models are displayed. The table shows clearly that complex models reduce the uncertainty with respect to alternative scenarios.

Figure 7 shows one of four possible scenarios for the development of reflexes of 'moon' inferred by the DWST model for the *Arbre* reference phylogeny. The scenario proposes a pattern in which the word form *yuè* 月 'moon' was replaced by the compound *yuèguāng* 月光 'moon-light' in all dialects except from the Mǐn subgroup. While this may well reflect a realistic scenario, we also find homoplastic (reoccurring) transitions, especially from *yuèguāng* 月光 to *yuèliàng* 月亮 'moon-shine' in the Wú subgroup. Homoplasy may point to lateral transfer events (List et al. 2014, Dagan and Martin 2007), but our knowledge regarding lexical evolution during the history of the Chinese dialects is still very limited. It is extremely difficult to tell with certainty whether the common reflexes of *yuèliàng* 月亮 in the Běijīng-Xiāng and the Wú subgroup reflect independent parallel developments or areal influence.

## 6. Conclusion

In this article, I have pointed to problems in the models used for phylogenetic reconstruction in linguistics resulting from a superficial treatment of historical relations between words. *Cognacy* is not a binary relation which is either present or not. Instead, we can distinguish different subtypes of cognacy, just as biologists can identify specific types of homology between genes. In an earlier paper, I proposed to compare the biological subtypes of homology (orthology, paralogy, xenology) directly with potential subtypes of historical word relations in linguistics (List 2014), but by concentrating on the major dimensions of lexical change proposed by Gévaudan (2007), namely morphological, semantic, and stratic change, I have shown that we can even go beyond the biological terminology and set up fine-grained schemas for historical relations in linguistics.

Which notion of cognacy we use for phylogenetic reconstruction crucially depends on the data we have at hand and the algorithms we intend to employ. I have shown that the inconsistencies in the treatment of historical relations between words have a direct impact on the way cognates are coded and data are analyzed in phylogenetic approaches. This was illustrated in detail for historical relations involving morphological change, especially compounding. If compounding is frequent and characteristic for a given language family, phylogenetic approaches which model lexical change merely as a process of cognate gain and cognate loss are inadequate and unrealistic. In order to take the different *degrees of cognacy* into account, I proposed to employ multistate instead of binary state models, and to further allow for potentially asymmetric transition tendencies among character states. The benefits of these models were demonstrated in a small experiment on semantic reconstruction applied to a lexicostatistical dataset of 23 Chinese dialect varieties. The results of this experiment strongly suggest that multistate models with asymmetric transition tendencies are superior to binary state models. What I have presented is, however, but a small step toward improved models of lexical change. More experiments including more language families need to be carried out. Instead of ancestral state reconstruction, we need to test the potential of multistate models for phylogenetic reconstruction in general. Probabilistic models, be they based on Maximum Likelihood (Felsenstein 1981) or Bayesian inference (Huelsenbeck et al. 2001), may prove really useful in this regard. In parsimony, we need to provide exact models for the transition between characters, and we always run the danger of overfitting our step matrices on a given dataset. Probabilistic models can help to estimate transition probabilities and could thus even provide new insights which go beyond cognacy and help us to detect major trends in lexical evolution, including morphological, semantic, and stratic change. In order to allow for these improved models of lexical change, however, we need to rethink the way we handle cognacy in our databases and start being more explicit in our annotations.

## Supplementary data

The most recent release of the accompanying software application can be found at https://zenodo.org/badge/latestdoi/5137/digling/beyond-cognacy-paper. An interactive application showing all inferred evolutionary scenarios for the Arbre phylogeny by Sagart (2011) is available at http://digling.github.io/beyond-cognacy-paper/.

## Acknowledgements

I thank three anonymous reviewers for challenging critics and helpful advice, and Gerhard Jäger for helpful comments made on an earlier version of this article. I am very grateful to Hans Geisler, who originally pointed me to many of the examples on lexical change which were treated in this article, to David Morrison, for important comments on the article and numerous fruitful discussions on questions of homology and cognacy, and to Laurent Sagart, who shared his ideas and his profound knowledge of Chinese dialect classification with me.

## Funding

This research was supported by the DFG research fellowship grant 'Vertical and lateral aspects of Chinese dialect history' (Grant No. 261553824), which is gratefully acknowledged.

## References

Arapov, M. V. and Xerc, M. M. (1974) *Matematičeskie metody v istoričeskoj lingvistike* [Mathematical Methods in Historical Linguistics]. Moscow: Nauka.

Atkinson, Q. D. and Gray, R. D. (2006) 'How Old is the Indo-European Language Family? Illumination or More Moths to the Flame?', in P. Forster and C. Renfrew (eds.) *Phylogenetic Methods and the Prehistory of Languages*, pp. 91–109. Cambridge and Oxford and Oakville: McDonald Institute for Archaeological Research.

Barðdal, J. and Eythórsson, T. (2012) 'Reconstructing Syntax: Construction Grammar and the Comparative Method', in H. Boas and I. A. Sag (eds.) *Sign-based Construction Grammar*, pp. 257–308. Stanford: CSLI Publications

Baxter, W. H. (1992) *A Handbook of Old Chinese Phonology*. Berlin: de Gruyter.

—— and Sagart, L. (2014) *Old Chinese. A New Reconstruction*. Oxford: Oxford University Press.

Ben Hamed, M. and Wang, F. (2006) 'Stuck in the Forest: Trees, Networks and Chinese Dialects', *Diachronica*, 23: 29–60.

Bohl, E. and Lancaster, P. (2003) 'Irreversible Markov Processes for Phylogenetic Models', *Numerical Linear Algebra with Applications*, 10: 577–93.

Bouckaert, R., et al. (2012) 'Mapping the Origins and Expansion of the Indo-European Language Family', *Science*, 337: 957–60.

Branner, D. P. (2000) *Problems in Comparative Chinese Dialectology. The Classification of Miin and Hakka*. Berlin and New York: Mouton de Gruyter.

Campbell, L. and Harris, A. C. (2002) 'Syntactic Reconstruction and Demythologizing 'Myths and the Prehistory of Grammars', *Journal of Linguistics* 38: 599–618.

Chung, K. S., Hill, N. W. and Sun, J. T.-S. (2014) 'Sino-Tibetan', in R. Lieber and P. Štekauer (eds.) *The Oxford Handbook of Derivational Morphology*, pp. 619–50. Oxford: Oxford University Press.

Coseriu, E. (1973) *Probleme der strukturellen Semantik* [Problems of Structural Semantics]. Tübingen: Narr.

Croft, W. (2008) 'Evolutionary Linguistics', *Annual Review of Anthropology*, 37: 219–34.

Dagan, T. and Martin, W. (2007) 'Ancestral Genome Sizes Specify the Minimum Rate of Lateral Gene Transfer During Prokaryote Evolution', *Proceedings of the National Academy of Sciences*, 104: 870–75.

de Saussure, F. (1916) *Cours de linguistique générale* [Course in General Linguistics]. Lausanne: Payot.

Dyen, I., Kruskal, J. B. and Black, P. (1992) 'An Indoeuropean Classification', *Transactions of the American Philosophical Society*, 82: iii–132.

Felsenstein, J. (1981) 'Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach', *Journal of Molecular Evolution*, 17: 368–76.

Fitch, W. M. (1970) 'Distinguishing Homologous from Analogous Proteins', *Systematic Zoology*, 19: 99–113.

——. (1971) 'Toward 'Defining the Course of Evolution: Minimum Change for a Specific Tree Topology', *Systematic Biology*, 20: 406–16.

——. (2000) 'Homology. A Personal View on Some of the Problems', *Trends in Genetics*, 16: 227–31.

Fox, A. (1995) *Linguistic Reconstruction*. Oxford: Oxford University Press.

Freeman, V. J. (1951) 'Studies on the Virulence of Bacteriophage-infected Strains of Corynebacterium Diphtheriae', *Journal of Bacteriology*, 61: 675–88.

Gates, J. P. (2012) 'Situ in Situ. Towards a Dialectology of Jiāróng (rGyalrong)', PhD thesis, Trinity Western University.

Geisler, H. and List, J.-M. (2013) 'Do Languages Grow on Trees? The Tree Metaphor in the History of Linguistics', in H. Fangerau, H. Geisler, T. Halling and W. Martin (eds.) *Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization*, pp. 111–24. Stuttgart: Franz Steiner Verlag.

Gévaudan, P. (2007) *Typologie des lexikalischen wandels* [Typology of Lexical Change]. Tübingen: Stauffenburg.

Gray, G. S. and Fitch, W. M. (1983) 'Evolution of Antibiotic Resistance Genes', *Molecular Biology and Evolution*, 1: 57–66.

Gray, R. D. (2005) 'Evolution: Pushing the Time Barrier in the Quest for Language Roots', *Science*, 309: 2007–8.

Gray, R. D. and Atkinson, Q. D. (2003) 'Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin', *Nature*, 426: 435–9.

Greenhill, S. J. (2015) 'TransNewGuinea.org: An Online Database of New Guinea Languages', *PLoS One*, 10: e0141563.

Greenhill, S. J., Blust, R. and Gray, R. D. (2008) 'The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics', *Evol. Bioinformatics*, 4: 271–83.

Hamming, R. W. (1950) 'Error Detection and Error Detection Codes', *Bell System Technical Journal*, 29: 147–60.

Handel, Z. (2010) 'Old Chinese and Min', *Chûgoku Gogaku* 中國語學 [Bulletin of the Chinese Language Society of Japan], 257: 34–68.

Haspelmath, M. (2004) 'On Directionality in Language Change with Particular Reference to Grammaticalization', in O. Fischer, M. Norde and H. Perridon (eds.) *Up and Down the Cline – The Nature of Grammaticalization*, pp. 17–44. Amsterdam and Philadelphia: John Benjamins.

Hattori, S. (1961) 'A Glottochronological Study on Three Okinawan Dialects', *International Journal of American Linguistics*, 27: 52–62.

Hóu, J. (2004) *Xiàndài Hànyǔ fāngyán yīnkù* 現代漢語方言音庫 [Phonological Database of Chinese Dialects]. Shànghǎi: Shànghǎi Jiàoyù.

Huelsenbeck, J. P., Bollback, J. P. and Levine, A. M. (2002) 'Inferring the Root of a Phylogenetic Tree', *Systems Biology*, 51: 32–43.

——, et al. (2001) 'Bayesian Inference of Phylogeny and its Impact on Evolutionary Biology', *Science*, 294: 2310–14.

Jacques, G. (2016) 'On the Directionality of Analogy in a Dhegiha Paradigm', *International Journal of American Linguistics*, 82: 239–48.

Jensen, R. A. (2001) 'Orthologs and Paralogs – We Need to get it Right', *Genome Biology*, 2: 1002.1–1002.3.

Karlgren, B. (1954) 'Compendium of Phonetics in Ancient and Archaic Chinese', *Bulletin of the Museum of Far Eastern Antiquities*, 26: 211–367.

Kassian, A., Zhivlov, M. and Starostin, G. S. (2015) 'Proto-Indo-European-Uralic Comparison from the Probabilistic Point of View', *The Journal of Indo-European Studies*, 43: 301–47.

Katičić, R. (1966) 'Modellbegriffe in der vergleichenden Sprachwissenschaft *[The Conception of Models in Historical Linguistics]*', *Kratylos*, 11: 49–67.

Kluge, F. and Seebold, E. (2002) *Etymologisches Wörterbuch der deutschen Sprache* [Etymological Dictionary of German]. 24 edn. Berlin: de Gruyter.

Koch, H. (1996) Reconstruction in Morphology, in M. Durie, (ed.) *The Comparative Method Reviewed. Regularity and Irregularity in Language Change*, pp. 218–63. New York: Oxford University Press.

Koonin, E. V. (2001) 'An Apology for Orthologs – or Brave New Memes', *Genome Biology*, 2: 1005.1–1005.2

Koonin, E. V. (2005) 'Orthologs, Paralogs, and Evolutionary Genomics', *Annual Review of Genetics*, 39: 309–38.

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

List, J-M. (2014) *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

List, J.-M.. (2015) 'Network Perspectives on Chinese Dialect History', *Bulletin of Chinese Linguistics*, 8: 42–67.

——, Cysouw, M. and Forkel, R. (2016) *Concepticon: A Resource for the Linking of Concept Lists*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://concepticon.clld.org>.

—— and Moran, S. (2013) 'An Open Source Toolkit for Quantitative Historical Linguistics', in *Proceedings of the ACL 2013 System Demonstrations*, pp. 13–18. Stroudsberg: Association of Computational Linguistics

—— et al. (2014) 'Networks of Lexical Borrowing and Lateral Gene Transfer in Language and Genome Evolution', *Bioessays*, 36: 141–50.

Matisoff, J. A. (2000) 'On the Uselessness of Glottochronology for the Subgrouping of Tibeto-Burman', in C. Renfrew, A. McMahon and L. Trask (eds.), *Time Depth in Historical Linguistics*, pp. 333–71. Cambridge: McDonald Institute for Archaeological Research.

Meiser, G. (1998) *Historische Laut- und Formenlehre der lateinischen Sprache* [Historical Studies of the Sounds and the Forms of Latin]. Darmstadt: Wissenschaftliche Buchgesellschaft.

Meyer-Lübke, W. (1911) *Romanisches etymologisches Wörterbuch* [Etymological Dictionary of Romance]. Heidelberg: Winter.

Michael, L., et al. (2015) 'A Bayesian Phylogenetic Classification of Tupí-Guaraní', *LIAMES*, 15: 193–221.

Morrison, D. (2015) 'Molecular Homology and Multiple-Sequence Alignment: An Analysis of Concepts and Practice', *Australian Systematic Botany*, 28: 46–62.

Needleman, S. B. and Wunsch, C. D. (1970) 'A Gene Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins', *Journal of Molecular Biology*, 48: 443–53.

Nelson-Sathi, S., et al. (2013) 'Reconstructing the Lateral Component of Language History and Genome Evolution using Network Approaches', in H. Fangerau, H. Geisler, T. Halling and W. Martin (eds.), *Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization*, pp. 163–80. Stuttgart: Franz Steiner Verlag.

Norman, J. (1988) *Chinese*. Cambridge: Cambridge University Press.

—— and Coblin, W. S. (1995) 'A New Approach to Chinese Historical Linguistics', *Journal of the American Oriental Society*, 115: 576–84.

Nunn, C. L. (2011) *The Comparative Approach in Evolutionary Anthropology and Biology*. Chicago and London: University of Chicago Press.

Oesterreicher, W. (2001) 'Historizität, Sprachvariation, Sprachverschiedenheit, Sprachwandel /Historicity, Language Variation, Language Diversity, Language Change/', in M. Haspelmath (ed.) *Language Typology and Language Universals*, pp. 1554–95. Berlin and New York: Walter de Gruyter.

Owen, R. (1843) *Lectures on Comparative Anatomy*. London: Longman, Brown, Green, and Longmans.

Pagel, M. (2009) 'Human Language as a Culturally Transmitted Replicator', *Nature Reviews Genetics*, 10: 405–15.

Pagel, M. D. (1999) 'Inferring the Historical Patterns of Biological Evolution', *Nature*, 401: 877–84.

Petsko, G. A. (2001) 'Homologuephobia', *Genome Biology*, 2: 1002.1–1002.2.

Pfeifer, W. (1993) *Etymologisches Wörterbuch des Deutschen* [Etymological Dictionary of German], 2 edn. Berlin: Akademie. <http://www.dwds.de/>.

Pulleyblank, E. (1984) *Middle Chinese: A Study in Historical Phonology*. Vancouver: UBC Press.

Ringe, D., Warnow, T. and Taylor, A. (2002) 'Indo-European and Computational Cladistics', *Transactions of the Philological Society*, 100: 59–129.

Rix, H. et al. (2001) *Lexikon der Indogermanischen Verben* [Lexicon of Indo-European Verbs], Wiesbaden: Reichert.

Sagart, L. (2011) 'Classifying Chinese Dialects / Sinitic Languages on Shared Innovations', Paper, presented at the Séminaire Sino-Tibétain du CRLAO. <https://www.academia.edu/19534510/Chinese_dialects_classified_on_shared_innovations>.

Sankoff, D. (1975) 'Minimal Mutation Trees of Sequences', *SIAM Journal on Applied Mathematics*, 28: 35–42.

Satterthwaite-Phillips, D. (2011) 'Phylogenetic Inference of the Tibeto-Burman Languages or on the Usefuseful of Lexicostatistics (and "megalo"-comparison) for the Subgrouping of Tibeto-Burman', PhD thesis, Stanford University, Stanford.

Sonnhammer, E. L. and Koonin, E. V. (2002) 'Orthology, Paralogy and Proposed Classification for Paralog Subtypes', *Trends in Genetics*, 18: 619–20.

Starostin, G. S. (2013) 'Lexicostatistics as a Basis for Language Classification', in H. Fangerau, H. Geisler, T. Halling and W. Martin (eds.), *Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization*, pp. 125–46. Stuttgart: Franz Steiner Verlag.

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

——. (1955) 'Towards Greater Accuracy in Lexicostatistic Dating', *International Journal of American Linguistics*, 21: 121–37.

Syvanen, M. (1985) 'Cross-species Gene Transfer. Implications for a New Theory of Evolution', *Journal of Theoretical Biology*, 112: 333–43.

Taylor, J. S. and Raes, J. (2004) 'Duplication and Divergence: The Evolution of New Genes and Old Ideas', *Annual Review of Genetics*, 38.

Ternes, E. (1987) *Einführung in die Phonologie* [Introduction to Phonology]. Darmstadt: Wissenschaftliche Buchgesellschaft.

Trask, R. L. (2000) *The Dictionary of Historical and Comparative Linguistics*. Edinburgh: Edinburgh University Press.

Vaan, M., (ed.) (2008) *Etymological Dictionary of Latin and the Other Italic Languages*. Leiden and Boston: Brill.

Walkden, G. (2013) 'The Correspondence Problem in Syntactic Reconstruction', *Diachronica*, 30: 95–122.

Wang, W. S-Y. (1997) 'Languages or Dialects?', *The CUHK Journal of Humanities*, 1: 54–62.

Weinreich, U. (1974) *Languages in Contact*, 8th edn. The Hague and Paris: Mouton.

Wilkins, D. P. (1996) 'Natural Tendencies of Semantic Change and the Search for Cognates', in M. Durie (ed.), *The Comparative Method Reviewed. Regularity and Irregularity in Language Change*, pp. 264–304. New York: Oxford University Press.

Wodtko, D., Irslinger, B. and Schneider, C. (2008) *Nomina im Indogermanischen Lexikon* [Nouns in the Indo-European Lexicon]. Heidelberg: Winter.

Wurzel, W. U. (1985) 'Morphologische Natürlichkeit und morphologischer Wandel. Zur Vorhersagbarkeit von Sprachveränderungen [Morphologigal Naturalness and Morphological Change. On the Predictability of Language Change]', in J. Fisiak (ed.) *Papers from the 6th International Conference on Historical Linguistics*, pp. 587–99. Amsterdam: John Benjamins.

Yóu, R. (1992) *Hànyǔ fāngyánxué dǎolùn* 漢語方言學導論 [Chinese Dialectology]. Shànghǎi: Shànghǎi Jiàoyù

Zhang, J. (2003) 'Evolution by Gene Duplication: an Update', *Trends in Ecology and Evolution*, 18.