Papers in Historical Phonology

http://journals.ed.ac.uk/pihph
ISSN 2399-6714

Volume 7, 44-80
DOI: 10.2218/pihph.7.2022.7405

What are cognates?

Marieke Meelen
University of Cambridge

Nathan W. Hill
Trinity College Dublin

Hannes Fellner
University of Vienna

Abstract

The popularity of computational methods in historical linguistics has primarily been motivated by mere access to the new methods themselves, rather than by looking for tools to solve problems. Investigators have looked for problems with which to showcase their tools. This dynamic is one reason why eye-catching but long-solved problems, such as the homeland of the Indo-Europeans (Gray & Atkinson 2003) have received more attention than genuinely unsolved or controversial questions, such as how to incorporate the Hittite *ḫi*-conjugation into an understanding of the Indo-European verbal system (Jasanoff 2003). One assumption of Bayesian methods is that cognacy can be conceptualized as binary. Although this is how historical linguists themselves often speak, it is not how they work. The goal of this article is to more precisely delimit what is meant when we call two words cognate, to emphasize that this is not a binary relation, but to suggest that this relationship can still be modeled formally.

1 Introduction

The idea of 'cognates' is fundamental to research in historical linguistics, both that carried out in a traditional framework and that making use of recent computational methodologies (cf. Labat & Lefever 2019). The term 'cognate' can be used both for languages and for linguistic material, usually words. According to Crystal (2008, 83), for instance, a cognate is a 'language or a linguistic form which is historically derived from the same source as another language/form'. This definition, along with similar formulations (e.g. Trask 2015, von Mengden 2008, Bynon 1977 etc.) raises many questions. To say 'historically derived' assumes a method, i.e. the 'Comparative Method', to reconstruct the sounds of the forms in and the overall phonological system of the proto-language, from which the cognate derives. Nonetheless, systematic, implementable accounts of the Comparative Method are scarce; those coming close generally start with 'compare words from cognate languages that have similar meanings'. In both the traditional and computational contexts the means of deciding whether two words are cognates or not remains largely opaque; there are no clear and explicit heuristics for determining the cognacy of two words. Basic textbooks teach via anecdotal examples (Anttila 1989, Campbell 2004, Hock 1991) and more advanced methodological works have a conceptual or theoretical focus that is not aimed at providing a practically implementable formalization of methodology as a series of steps (Hale 2007, Hoenigswald 1960). In our view, one reason that the historical linguistics of non-Indo-European languages lags behind work on the Indo-European family is precisely because so much of Indo-European practice remains tacit, implicitly absorbed as disciplinary norms, and consequently not communicated to those working elsewhere (Schwink 1994, 29, Fellner & Hill 2019a). Remarkably, in computational phylogenetic contexts, cognacy is typically left to human experts (e.g. Gray & Jordan 2000, Chang, Cathcart, Hall & Garret 2015).

To rigorously formalize the comparative method would yield two paramount benefits: (1) to better teach the method to practitioners, particularly those working outside of Indo-European, (2) to potentially automate certain stages of the workflow and thereby spare the time of researchers to concentrate on the conceptually more challenging steps. However, before we can formalize the workflow of the comparative method it is necessary to formalize what is meant by the key concepts that this workflow makes use of; this paper focuses on determining what is meant by the relationship of cognacy between two words or morphemes. We propose a systematic method for diagnosing cognates and a practical workflow that is easy to implement. We present:

- a workflow for establishing cognate sets (section 3)
- a typology of cognates and a hierarchy of cognacy (section 4)
- diagnostics for categorising cognates (section 5)

We begin by exploring the boundaries of what may legitimately be called cognates (section 2) by examining two pairs of words, the famous comparison of Greek θεός and Latin *deus*, a pair that looks related but are not (section 2.1), and the comparison of English *tooth* and Old Irish *dofúaid* 'he has eaten', words distant in form and not close in meaning but that descend from the same Indo-European root (section 2.2). Next we discuss the value of clear diagnostics in other areas of linguistics, demonstrating that there are currently no diagnostics to determine the limits of cognacy (section 2.3). The bulk of the paper establishes a typology of cognacy (section 4), dividing individual cases among strict (section 4.1), medium (section 4.2), and weak (section 4.3). The subsequent section discusses the challenges to explicitly diagnosing the different types of cognacy, also proposing some solutions (section 5). We end the paper with some preliminary conclusions, including directions for future research (section 6).

2 The limits of cognacy

In this section we explore the boundaries of what may be called cognates by looking at two examples. The main heuristic here is the phonological *form*, measured by the regularity of sound laws: words or morphemes are cognates if, and *only* if, their phonology can be reconstructed following regular sound changes back to the proto-language (or whichever intermediate stage at which two cognate languages started to diverge). In addition, the *meaning* and *function* of the two forms should be the same or at least similar. This second heuristic is less black-and-white and will be discussed in more detail in sections 5.2 and 5.3 below.

The first example is the famous comparison of Greek θεός and Latin *deus*, which looks obvious but turns out to be false (section 2.1). The second, the comparison of English *tooth* and Old Irish *dofúaid* 'he has eaten' is *prima facie* ridiculous but turns out to be genuine (section 2.2). Consideration of these two cases sheds light on the fact that 'cognacy' is not once and for all, but instead that words come in and out of cognacy as scientific understanding deepens.

2.1 Greek θεός 'god' and Latin *deus* 'god': worse than it looks

The comparison of Latin *deus* and Greek θεός, so familiar from the handbooks (Fortson 2010, 25), is the *Paradebeispiel* for the methodological principle that form trumps meaning in etymology. The two words look similar and mean the same thing, but an initial Latin *d*- should correspond to Greek δ- (e.g. Lat. *domus* 'house' and Gk. δόμος 'house') and not to θ-. The comparison of θεός and *deus* dates to before Franz Bopp (1791-1867); Bopp's student, August Friedrich Pott included both *deus* and θεός as cognates of Sanskrit *deva* 'god' in his *Etymologische Forschungen auf dem Gebiete der indogermanischen Sprachen* (1833-6), but noted the phonological irregularity (Davies 1998, 173-174). Scholars such as Theodor Benfey

(1837) and Georg Curtius (1862) slowly brought the *opinio communis* to reject this proposal, with Max Müller holding on to the defunct comparison as late as 1875.

The power of the comparative method is to show that obvious looking cognates such as these are in fact impossible from the point of view of historical phonology. But with the benefit of hindsight it is easy to forget that in its day the comparison was not foolish, but necessary; in less well studied language families similarly plausible, but unjustifiable comparisons are rife. Consider from the Trans-Himalayan family the comparison of Old Tibetan *dmyig* 'eye' with Burmese *myak*, and Chinese 目 *mjuwk* < *C.muk. The vowel correspondence *i* : *a* : *u* is unique to this example, but to dismiss the cognacy of the items at the current state of research would be premature.

2.2 English *tooth* and Old Irish *dofúaid* 'he has eaten': better than it looks

At the opposite end of the conceptual spectrum from the 'god' example is when two words do have a shared history, but that the naïve proposal of their common ancestry would be unwarranted. The initial *t-* in English *tooth* and the final *-d* in Old Irish *dofúaid* 'he has eaten' both continue the _*d_ of the Indo-European root _*h_$_1$_ed-_ 'eat'. On the one hand, Old Irish *dofúaid* 'he has eaten' is a suppletive third person singular perfect deuterotonic stem (see Thurneysen 1946, 27-29, 351-352) of *ithid* 'eat'. A transponant of *dofúaid* for Proto-Insular-Celtic would be _*dī-wo-ād_, in turn from Proto-Celtic _*dē-uɸo-ād-e_, and, if ultimately projected back into Indo-European, _*dē_ + _*upo_ + _*h_$_1$_e-h_$_1$_od-e_, two preverbs before a third person singular perfect. The form _*h_$_1$_e-h_$_1$_od-e_ is itself straightforwardly the singular perfect of the root _*h_$_1$_ed_. On the other hand, English *tooth* is from Old English *tōþ*, from Proto-Germanic *_tanþs_ (cf. Old Saxon *tand*, Dutch *tand*, Go. *tunþus*), from Proto-Indo-European _*h_$_1$_dónts_; cognates include Latin *dent-*, Aeolic Greek ἔδοντ- (see Ringe 2006, 70), Old Irish *dét*, and Lithuanian *dantìs*. The stem _*h_$_1$_dónt-_ is itself straightforwardly the active participle of the root _*h_$_1$_ed_ 'eat'. Nonetheless, it is not in accord with normal practice in historical linguistics to regard *tooth* and *dofúaid* as cognate. Permitting such examples would countenance all of the excesses of the long rangers. From the ontological perspective such cases are clearly cognate, but from the epistemological perspective this loose a notion of cognacy has little practical methodological value, unless we define a clear hierarchical typology of cognacy (for this see section 4). The knowledge that these words are in any sense related is the end product of a vast amount of research, it is not the starting point for an investigation.

2.3 The benefits of clear diagnostics

Like in other sciences, a linguistic study often starts with an observation and an attempt to accurately describe the object of study: a sound, word, sentence etc. in one or more language(s). Ideally we then move beyond the description to try and explain how and why we observe the patterns, constructions, forms in the way we do. In the introduction to a collection of papers *Diagnosing Syntax*, Cheng & Corver (2013) compare the study of syntax and the discovery of 'underlying' or 'hidden' structures to the work of physicians: the nature of the illness or disorder can be identified based on a patient's signs and symptoms. Similarly, a careful and rigorous study of the properties and characteristics (i.e. the symptoms) of a syntactic phenomenon, for example, can identify it: 'the nature of an object or phenomenon is understood by means of the ability to discern relevant features of that object or phenomenon' (Cheng & Corver 2013, 1). In order to diagnose a physical condition, physicians conduct a range of tests or diagnostic procedures (e.g. blood or urine tests). A good syntactician should thus be a good diagnostician: capable of designing and consistently conducting the right tests to identify a phenomenon within a language or even cross-linguistically. Within a language, these tests can be quite specific. In a language like Dutch for example one can distinguish unaccusative and unergative intransitive verbs[1] by testing which auxiliary they take in the perfect, i.e. 'be' or 'have' respectively:

(1) a. Hij is vertrokken.
he be.3SG departed
'He has departed.' ('be' AUX: unaccusative)
b. Hij heeft gedanst.
he have.3SG danced
'He has danced.' ('have' AUX: unergative)

This, along with a number of other tests, helps identify the type of intransitivity of certain verbs in Dutch (and some other languages, e.g. German,

[1] Unaccusative verbs are intransitive verbs whose subject is not considered to be the 'semantic agent' or 'external argument' (in generative grammar). This subject is therefore structurally and semantically similar to the direct object or patient of a transitive verb, e.g. *arrive*, *die*. They contrast with unergative intransitive verbs whose subject voluntarily initiate the action, e.g. *dance*, *run*. See Alexiadou, Anagnostopoulou & Everaert (2004).

French, Italian, etc.). However, as is clear from the translation, the same test will not work for English, because Present-Day English only has one auxiliary for the perfect ('have'). To diagnose unaccusative verbs in English, other tests are needed, such as nominal modifiers or resultative adjuncts. To give an example of the former, past participles of unaccusative verbs in English can be used as active nominal modifiers, whereas those of unergatives cannot:

(2) a. The departed guests. / The melted snow. (OK: unaccusative)
b. *The danced girl. / *The slept child. (NOT OK: unergative)

Like syntacticians, historical linguists, when looking for answers to 'how' and 'why' questions, are confronted with 'hidden structures'. Phonological comparison and reconstruction are good examples of such hidden structures. Our objective in this paper is to identify the nature of the phenomenon, in this case, the level of inherited similarity or 'cognacy'. In order to do that, as historical linguists, we also need a set of clearly defined tests. A lack of clear heuristics and diagnostics makes it difficult to verify and compare results consistently. Just like theoretical syntacticians or linguists in other subdisciplines like neuro- and psycholinguistics, historical linguistics would benefit from more well-defined ways to make predictions and to test results. If we were to say with greater precision what is meant with the claim that two words are 'cognate' and provide clear methods for identifying whether two forms are cognate, this would help those areas of historical linguistics where progress is currently stymied.

3 Workflow for establishing cognate sets

The aforementioned two pairs of examples illustrating the limits of cognacy (section 2.1 and section 2.2) highlight a distinction between what we can conveniently call *comparanda*, words suspected of descent from a single form (Latin *deus* and Greek θεός), and *comparata*, words that have been shown to descend from a single form (English *tooth* and Old Irish *dofúaid*). A suspicion can be more or less strong, a demonstration more or less secure, as such both suspicion and demonstration are scalar rather than binary predicates. Therefore, being a *comparandum* or a *comparatum* is a concomitantly complex affair. The best-behaved cognates are those where *ex ante* any observer would have a strong suspicion of their shared origin and their shared origin has been demonstrated in an *ex post facto* straightforward and watertight way, i.e. where a good *comparandum* is a good *comparatum*. As etymological research progresses one relies less on *comparanda* and more on *comparata*; the machinery of known historical phenomena become more powerful as they become more finely stated. The exactness of science replaces the groping of guess work. The rest of this section attempts to answer the question of how we change *comparanda* into a *comparata* at a specific moment in the history of research.

3.1 Step 1: Heuristics for finding comparanda

There are three elements essential to establishing cognate sets: a set of cognate languages, a set of potential cognates (*comparanda*) and, ideally, a body of existing knowledge to test against, namely, a set of sound correspondences (C) thought to be regular at a particular point in time (t); we call this set of regular sound correspondences C_t. For languages with well-established phylogenies and a large body of secondary literature, all three necessary elements are readily available. The existence of a well-developed set of sound correspondences (C_t), in particular, permits one to go straight to the diagnostics that help determine the type and level of cognacy (see section 5).

For under-researched languages families for which no such literature and resources exist, we need initial heuristics to get the workflow started. In such cases, the three essential elements may instead be conceptualized as steps in a preliminary mini-workflow:

1. Choose languages to compare
2. Choose words to compare
3. Choose a set of allowable correspondences (C_t)

The first step of the mini-workflow is not of methodological interest, since in principle all of the world's languages could be compared pairwise. In practice, languages will be compared to languages of presumed genetic affiliation or geographic proximity. As for the second step, existing computational methods for finding potential cognates may not yet be well-equipped to diagnose the type and level of cognacy, but they certainly permit the identification of potential *comparanda*, *faute de mieux*. LexStat is a prominent example of such an automatic cognate detection algorithm (List 2012); when computation time is costly, because words from many languages are compared, BipSkip is an alternative that performs faster, but less well (Rama & List 2019).

The third step is more difficult, when investigating *comparanda* from two languages that have never been looked at together before, no set of established sound correspondences exists. As a further heuristic in these cases, one can look to well-established sound correspondence sets in other language families to identify plausible candidate correspondences. The phoneme */t/* for instance, often corresponds with */t/, /d/* or /t^h/, but a correspondence between */t/* and a vowel or approximant, or even other stops like */p/* or */k/*, is unprecedented, or at least rare. Since in this case C_t is the allowed correspondence patterns at the very beginning of research ($t = 0$), we refer to this set of correspondence patterns as C_0. The heuristic under discussion populates C_0 with correspondence patterns that are widespread across the world's languages. To give a simple example, correspondences of identity such as {*m*, *m*, *m*}, {*n*, *n*, *n*}, {*h*, *h*, *h*} and {*s*, *s*, *s*}, one will certainly want to include in C_0 at this point. This initial step of populating C_0 results in a set of hypotheses only, which is important to bear in mind when phonological segments are aligned and morphemes are tested in the next step of the main workflow.

3.2 Step 2: Aligning and checking phonological segments

Once we have a set of *comparanda* and at least a start on a set of sound correspondences, we continue with the alignment of the phonological segments. It is important to note that although this may seem trivial to a trained historical linguist, this is a non-trivial task for a computer when the length of the phonological segments of the *comparanda* differs; particular difficulties are, for example, knowing when to permit a segment to compare to zero and whether to compare a diphthong with a vowel, or a sequence of vowel and glide (List 2014 and List, Walworth, Greenhill, Tresoldi & Forkel 2018). At this step we refine the 'historically derived from the same source' part of the initial definition of cognates we cited in section 1. In order to develop a straightforwardly implementable method, we propose to re-define the 'historical derivation' in terms of minimum requirements of cognacy: for *comparanda* to be cognate, at the very least they need to have one aligned phonological segment that can be found in the set of established sound correspondences between the respective languages. In addition, at a minimum it is necessary to tell an informal, intuitively plausible story about how one single meaning can develop into those seen in the *comparanda*.

After aligning all phonological segments, we check whether the resulting sound correspondences exist in our set of established correspondences (C_t) for the languages in question.[2] If the phonological segments of a root

[2] In keeping with the adage falsely attributed to Voltaire that etymology is 'où la voyelle ne fait rien, et la consonne fort peu de chose' ('where vowels count for nothing, and conson-

morpheme can be aligned and the resulting correspondences exist in C_t, the *comparanda* pass this initial test and can be called 'cognate'. If the concept of 'roots' in unclear in the languages under investigation, then this test can be relaxed to apply to any morpheme. If there is no single morpheme for which phonological segments can be aligned, the *comparanda* are rejected as cognates at this stage (see figure 2). Note that this immediately rules out the cognacy of the roots of Latin *deus* and Greek θεός, although it leaves open the possibility of considering the endings *-os* and *-us* as potential cognate candidates, which is the desired result. It does not *a priori* rule out the cognacy of English *tooth* and Old Irish *dofúaid* 'he has eaten', but the difficulty of aligning the one corresponding phonological segment immediately reveals the weakness of the cognacy — again, the desired result.

For newly compared languages, C_0 is pre-populated only with sound correspondences well attested across the globe. There may well be valid correspondences that have not yet been countenanced by any existing research tradition. In this case the *comparanda* that evince correspondence patterns not found in C_0 should not be rejected out of hand. Instead, one uses the *comparanda* to add to and verify the set of sound correspondences. This process of identifying correspondence patterns makes up the backbone of the comparative method and for this reason alignment is particularly essential in the early stage of research (Anttila 1972, 230, Koch 1996, 221, Dimmendaal 2011, 13, Weiss 2014, 128, Trask 2015, 196).

3.3 Step 3: Diagnosing cognacy type and level

Once we have established that we are in fact dealing with cognates, we can establish the type and level of cognacy. We present a number of diagnostic tests that evaluate the form (phonology), meaning (semantics) and function (morpho-syntax and pragmatics) of the *comparanda*. These diagnostics will first of all determine the level of cognacy ranging from strong ('strict cognates') to weak. Second, the diagnostics establish whether cognates are synonymous or non-synonymous in meaning and function. Among any level of cognates we can distinguish 'synonymous cognates' (Koch & Hercus 2013, 34), for those comparisons where both members maintain the inherited meaning unchanged, and 'non-synonymous cognates' for those comparisons where one or both members of the comparison have undergone semantic change. Strict synonymous cognates such

ants for very little'), one could assign greater weight to correspondences of consonants than those of vowels.

as German *Herz* 'heart' and English *heart* and German *Distle* 'thistle' and English *thistle*, etc. are the most straightforward type and play a unique role in the early days of research on a particular family (List 2019). It is no coincidence that automatic cognate detection algorithms, such as LexStat, require synonymous *comparanda* as their input.[3] As for the strict non-synonymous cognates, we can allow for major semantic changes along the lines of German *Zimmer* 'room' and English *timber*, or German *Zaun* 'fence' and English *town*, but what we cannot permit is complete semantic laxness.

And finally, diagnostics can classify *comparanda* belonging to certain sub-types, e.g. 'partial cognates', 'oblique cognates' or 'quasi-strict cognates'. These types and levels are treated in detail in section 4, and the diagnostics in section 5. After the application of these diagnostics, new cognates are labeled and categorised, for instance as 'strict synonymous', 'weak synonymous' or 'quasi-strict non-synonymous' cognates; they are now *comparata*. Where relevant, their sound correspondences, including any additional features such as their phonological context or conditioning, can now be added to the set of established sound correspondences (C_t).

3.4 Step 4: Extending and refining the sound correspondence set

After newly established cognates or *comparata* emerge from Step 2 and are labeled in Step 3, the sound correspondence set is re-calibrated to reflect the knowledge gained (e.g. the addition of a set of features from new examples to an established correspondence pattern).

m	a	n	s	o:	n	h	ɛ	m	h	oː	r	h	ʉː	s	m	ʉː	s
m	æ	n	s	ʌ	n	h	əʊ	m	h	ɛ	r	h	aʊ	s	m	aʊ	s
m	a	n	s	o:	n	h	aɪ̯	m	h	aː	r	h	aʊ̯	s	m	aʊ̯	s

Figure 1: Aligned cognates from Swedish, English, and German apud Anttila 1972, 230. Orthographic forms are rewritten as broad phonemic IPA transcription.

Suppose that we were for the first time investigating the relationship among Germanic languages. Figure 1 shows a few aligned *comparanda*, namely words meaning 'man', 'son', 'home', 'hair', 'house', and 'mouse' in

[3] It is also not a coincidence that the words in figure 1 (discussed anon) are strict synonymous cognates.

modern Swedish, English, and German (following Anttila 1972, 230).These words exhibit the correspondences *m*:*m*:*m* (3x), *n*:*n*:*n* (2x), *h*:*h*:*h* (3x), *s*:*s*:*s* (3x), *ʉː*:*aʊ*:*aʊ* (2x). The first four correspondence patterns pass muster, since we included the correspondence patterns {*m*, *m*, *m*}, {*n*, *n*, *n*}, {*h*, *h*, *h*} and {*s*, *s*, *s*} *ex hypothesi* in C_0. Step 3 will have classified these comparisons as quasi-strict synonymous cognates, quasi-strict because the fifth correspondence pattern, *ʉː*:*aʊ*:*aʊ*, was not included in C_0. However, on the basis of these, and other, examples, the pattern can be added to C_1. In a future iteration of the workflow, these comparisons will come out as strict synonymous cognates.

The major conceptual hurdle is that it is not always clear how to distinguish the major sound correspondences relating two languages from those that should be regarded as one-offs. For example, when we consider the Latin word *quinque* 'five', and contemplate the reason for it lacking the *p- of its progenitor **pénk^{w}e* 'five' (Gk. πέντε, Skt. *páñca*, Lithuanian *penkì*), it is reasonable to see the non-alignable element as contamination from **k^{w}etu̯or-* (Gk. τέσσαρες, Lat. *quattuor*, Av. *caθβar*). However, the same form can be explained with a sound change **p* ... **k^{u}* > **k^{w}* ... *k^{w}*, whereby the **p-* is assimilated to the labiovelar of a following syllable (Weiss 2009, 73). The latter explanation has the advantage of also explaining Lat. *coquit* 'cooks' as the outcome of **pek^{w}-e-ti* 'cooks' (cf. Skt. *pacati*, Gk. πέσσω). In practice, what we can do is set an arbitrary frequency threshold to accept only the commonly attested patterns into C_t, e.g. those occurring in 15 or more cognates, but to lower the threshold as C_t becomes populated with more and more refined information about the historical phonology of the languages in question as the workflow goes through progressive iterations. Since the workflow itself will weed out spurious comparisons and add in non-obvious comparisons, the threshold chosen to accept new correspondence patterns after any given iteration really does not matter. To play it safe, the *most* common pattern not yet in C_t would be the only pattern examined, and if it is accepted, the whole workflow would be rerun.

Figure 2 presents a schematic overview of the entire workflow proposed in this paper.

4 The typology of cognates

In this section we present a typology of cognates, based on the level of similarity of their three core variables: form, meaning and function. In the previous section we have already established a crucial minimum level of similarity in form: two cognates must both contain at least one segment

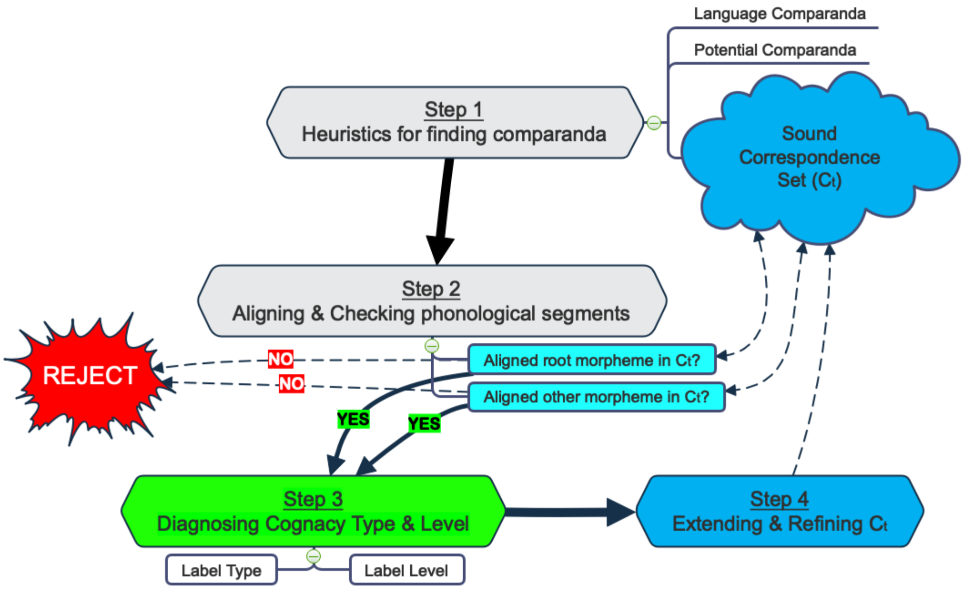
Language Comparanda
Potential Comparanda
Step 1
Heuristics for finding comparanda
Sound Correspondence Set (Ct)
Step 2
Aligning & Checking phonological segments
REJECT
NO
NO
Aligned root morpheme in Ct?
Aligned other morpheme in Ct?
YES
YES
Step 3
Diagnosing Cognacy Type & Level
Step 4
Extending & Refining Ct
Label Type
Label Level

Figure 2: Schematic workflow of establishing cognates.

that continues the same inherited segment in a common ancestral ‘root’ morpheme.[4]

The definition of ‘root’ morpheme can vary depending on the language family, but in principle cognacy can always be established (or rejected) on the level of the morpheme. Morphemes such as derivational suffixes, case endings, or verb inflections can thus also be compared and tested for cognacy (see note 6).

Prototypical examples of the main types of cognacy are discussed in three broad subsections ranging from phonologically strong to weak cognates. Alongside the phonological form, most definitions of cognacy refer to a similarity in meaning (semantics), to which we add a potential similarity in function (morpho-syntax and pragmatics). We adopt the notions of ‘synonymous’ and ‘non-synonymous’ cognates to reflect the latter variables and argue that the level of similarity in meaning and function can be established for each pair of cognates, no matter how strong or weak in phonological form.

[4] Admittedly Fr. *être* and Sp. *ser* are not cognate by this measure, but many forms of their paradigm are (e.g. Fr. *sommes* and Sp. *somos*). The correct theorization of such tricky cases is best addressed in future research.

4.1 Strict cognates

We propose the term *strict cognates* for those words or word parts for which we can demonstrate that their change was following the regular 'laws' of sound change. Such cases contrast with words whose histories include additional factors such as morphological derivations that impact directly on pronunciation, or sporadic sound changes due to analogy, assimilation, metathesis, etc.

The strict level of cognacy is only common at a fairly shallow time depth for a small selection of languages, since at a greater time depth erratic analogical developments more and more affect an ever greater portion of the vocabulary. After enough time such analogical development even compromise the ability for an analyst today to find the regular correspondences.

Afro-Asiatic is, for example, a family so ancient as to make the identification of regular phonological correspondences exceedingly difficult (Huehnergard 2004, 141); we face similar problems in the deep reconstruction of Trans-Himalayan (Hill 2019), and, as discussed for some examples in detail below, even in Indo-European linguistics, it is often difficult to find reflexes of well-established proto-forms which are regular in all respects. Adding further languages to the comparison asymptotically increases the number of exceptional analogical developments in a comparison (since 1 analogical innovation among *n* languages leads to n irregular comparisons so), strict cognates are much easier to identify in a pairwise fashion, two languages at the time. Strict cognates (cognates, such as the examples in figure 1, where all phonological segments can be aligned and for which the correspondences can be found in C_t, see section 4.1) have a unique importance for the discovery of correspondence patterns (List 2019).

4.2 Medium cognates

To classify some types of relationships as medium strength cognates is primarily an expository exercise, in other words 'medium cognates' are all those that are neither strong nor weak. We distinguish two types: 'quasi-strict cognates' (section 4.2.1) and 'word equations' (section 4.2.2). In this section we first discuss the three types of quasi-strict cognates and then we provide examples of word equations.

4.2.1 Quasi-strict cognates

In some cases two related words have mostly been affected by regular phonological change, but one or both of them have also been affected by a non-phonological change that has resulted in an exceptional status for an individual segment. Since such cases are almost as straightforward as strict cognates, we refer to them as ‘quasi-strict’. Nonetheless, the exceptional segment needs to be located and a reasonable explanation put forward to account for its existence. Quasi-strict cognates arise particularly due to three causes: paradigm internal analogy, contamination, and inter-dialectal borrowing.

Paradigm-internal analogy Some single segmental exceptions to regularity result from paradigm internal analogies. Because of the grammatical motivation for the analogy, such cases may have the appearance of grammatically conditioned sound changes. For example, Crowley finds that the failure of initial _*l_ loss in Paamese verbs, such as _loh_ ‘he runs’ ≪ _*oh_ is ‘a clear example of a sound change that does not involve purely phonological conditioning factors but also involves grammatical conditioning’ (2010, 173), but he fails to mention that it is only in non-negated third singular realis verb forms that *l- would have been word initial (Crowley 1982, 129–130). The paradigmatic pressure to analogically restore l- in this environment was overwhelming (e.g. _navō_ ‘I stink’ : _vō_ ‘he stinks’ :: _naloh_ ‘I run’ : X ‘he runs’, with _loh_ ‘he runs’ replacing inherited _*oh_) (see Hill 2014, 222).

Contamination In cases of _contamination_ (Hock 1991, 197–199, Trask 2000, 72–73, etc.) the pronunciation of a word is affected by the pronunciation of a word with which it is semantically associated (see esp. Hockett 1967). A well known instance is that of Indo-European _*k_^{w}_etu̯ṓr_ ‘four’ (cf. Skt. _catvā́r-_) irregularly becoming Proto-Germanic _*petu̯ṓr_ > _*fedu̯ōr_ (Go. _fidwōr_, OEng. _fēower_) under the influence of the _*p-_ in _*pénk_^{w}_e_ ‘five’ (Gk. πέντε, Skt. _páñca_, Lith. _penkì_). If we compare, for example Sanskrit _catvā́r-_ and Gothic _fidwōr_, some segments are alignable according to regular phonology (_-a-_:_-i-_, _-t_:_-d-_, _-r_:_-r_) but it would be a mistake to mechanically align the _f-_ of Gothic with Sanskrit _c-_, since that correspondence does not regularly recur in other vocabulary. To give another example, also well-known from textbooks (Hock 1991, 230, Trask 2015, 31, etc.), German _Bräutigam_ and Dutch _bruidegom_ ‘groom’ are strict cognates, but neither is alignable with English _bridegroom_, because the second _-r-_ of the latter has no reflex in the other two languages, resulting as it does from contamination with _groom_ and/or the existing _-r-_ in _bride_.

Interdialectal borrowing Borrowing between closely related languages can also lead to what are epistemologically (at least initially) indistinguishable from quasi-strict cognates, but which, as loans are by definition _not_ cognates. An example for this process is German _Damm_ 'dam' (< OHG _tamb_), which shows initial _d_ instead of expected _t_ due to contact with Low German varieties (where _d_ is the regular initial). Although different processes are at work here (close language contact and replacement of one item by a similar cognate item from a closely related language variant vs. language-internal modification of a form due to analogy with a form of a different meaning in the same language), the resulting patterns are very similar, in so far as strictness of sound correspondence patterns is maintained for the most part throughout the word, but one segment does not follow the expected pattern.

4.2.2 Word equations

In addition to these 'quasi-strict' cognates, there is another type of medium cognate which we label 'word equations'. These are cognates exhibiting the same form derived from the same root, continuing at least one (but not all) inherited grammatical feature(s). In Indo-European linguistics two forms enter a word equation when they exhibit the same form of the same root and continue some inherited grammatical feature under discussion (Vine 1993, 49, Jasanoff 2003, 3, 13, et passim, Clackson 2007, 187, 210, Weiss 2009, 430). Two of Jasanoff's (2003) equations give a feel for how the term is used. He supports the continuity of the Hittite _mi_-conjugation present singular 3rd person personal ending-_zi_ from the Proto-Indo-European primary active ending _*-ti_ with the following word equations (Jasanoff 2003, 3):

- Hitt. 3rd sing. _ēšzi_ = Vedic _ásti_ = Gk. ἔστι < PIE _*h$_1$és-ti_
- Hitt. _kuenzi_ 'slays' = Vedic _hánti_ < PIE _*g^{wh}én-ti_

Note that all strict cognates are necessarily word equations (but not all word equations are strict cognates),[5] although one would tend not to

[5] As an example of an imperfectly alignable word equation, Jasanoff equates Hitt. _mimma-_ 'refuse' (_prima facie_ from _*mimne-_) and Gk. μίμνω 'stand fast' (< _*mimn-_), while arguing that reduplicated presents with the reduplicating vowel -_i_- are associated with the the Hittie _ḫi_-conjugation (Jasanoff 2003, 129). The Hittite stem final vowel -_a_ is not alignable with anything in Greek; Jasanoff explains it as an analogical innovation on the model of the 3rd pl. _mimmanzi_ (2003, 131), i.e. _danzi_ 'they give' : _dāi_ 'he gives' :: _mimmanzi_

refer to a set of monomorphemic strict cognates as word equations. It is perhaps no surprise that in the research traditions of those language families, such as Austronesian or Trans-Himalayan, with members less rich in morphology — or with poorly understood morphology — the term ‘word equation’ does not appear.

4.3 Weak cognates

Weak cognates are morphologically altered with respect to the proto-form. We sub-categorize weak cognates into two types, namely partial cognates (section 4.3.1), and oblique cognates (section 4.3.2).

4.3.1 Partial cognates

We define ‘partial cognate’ as forms which contain at least one morpheme that is strictly cognate and at least one of the comparanda contains an additional morpheme not present in the other. Thus, Spanish _sol_ (< Vulgar Lat. _sōl_; Lat. _sōl_) and French _soleil_ (< Vulgar Lat. _*sōliculus_ REW, §8067) are related as partial cognates, as are Atsi _mau^{21}mjiŋ^{51}_ ‘thunder’ and Maru _mjaŋ^{31}k^{h}a^{35}_ ‘wolf’, since they both continue an inherited morpheme for ‘thunder’; the Maru word has the morphological structure ‘thunder’ + ‘dog’ (cf. Maru _lǎ^{31}k^{h}a^{5}_ ‘dog’) (Hill & List 2017, 68).

Among partial cognates, one can distinguish a subtype of ‘root cognates’ for cases when the two reflexes exhibit the same form of the same root (Trask 2000, 290).[6] For example, Latin _lātus_ ‘borne’ (< _*tl̥h_2-tó-_) and Old Church Slavonic _tĭla_ ‘foundation, bottom’ (< _*tl̥h_2-ó-_). Both words continue the same form of the same root (the zero-grade _*tl̥h_2-_), but also contain non-cognate concatanating morphology (the suffixes _*-tó-_ and _*-ó-_). The meaning of ‘root’ will of course depend on the specific morphological profiles of particular language families. For any language it is likely possible to work with a definition such as ‘inherited morpheme in a word that, of

‘they refuse’ : X = _mimmai_ ‘he refuses’. The two members of the word equation are not alignable but they continue the same form of the same stem (_*mimn-_) and share a relevant grammatical feature (here present reduplication with the reduplicating vowel _-i-_).

[6] A proto-language could have had a form Root_1+Suffix_1, where no daughter language preserves this combination, offering only Root_2+Suffix_1 or Root_1+Suffix_2. As such, there are partial cognates that are not root cognates, forms that share a stem but have a different root (e.g. 3sg.prs. _*g^{w}m̥_**_-sḱ_**_-é-ti_ > Skt. _gacchati_ and 1sg.prs. _*ǵi-ǵneh$_3$_**_-sḱ_**_-ó-h$_2$_ > Gk. γιγνώσκω ‘know’). Nonetheless, a word equation (e.g. 3sg.prs. _*g^{w}m̥-sḱ-é-ti_ > Skt. _gacchati_ ‘go’ and 1sg.prs. _*g^{w}m̥-sḱ-ó-h$_2$_ > Gk. βάσκω) is much better evidence for the reconstruction of a suffix.

the morphemes in the word, has the lowest synchronic frequency across the lexical entries of that language.'

4.3.2 Oblique cognates

As described above (section 2.2), cognates such as English _tooth_ and Old Irish _dofúaid_ are not conventionally called cognates because _prima facie_ there is no wisdom in bringing together these words for comparison. We next turn to comparisons that are much more fruitful, but perhaps no less complex in terms of their historical relatedness. Consider English 'feather' compared to Greek πτερόν 'feather, wing'. Indo-European had an original proterokinetic heteroclitic noun with rectus stem _*pét-r̥_ and obliquus stem _*pt-én-_ (cf. Hitt. _pettar_, _pettan-_ 'wing, feather'). English 'feather' derives from _*pét-r-eh$_2$-_ 'collection of feathers' with the _*-eh$_2$_ collective suffixed to the inherited rectus stem _*pét-r-_. In turn, Greek πτερόν continues _*pt-er-ó-_ 'feathery thing', a possessive _*-o-_ derivative of a stem _*pt-er-_, which is an analogically renewed obliquus stem, i.e. rectus _*pér-tu-_ (ON _fjǫrðr_) : obliquus _*pr̥-téu̯-_ (Lat. _portus_, Eng. _ford_) 'crossing' :: rectus _*pét-r_ : obliquus X, X = _*pt-ér-_, or the like. The comparison of 'feather' and πτερόν is what Trask uses in his definition of 'oblique cognate' (Trask 2000, 235). Trask defines an oblique cognate as '[t]wo or more words in related languages which continue alternate forms of a single root in the ancestral language' (2000, 234–5). This definition refers to 'a single root', so oblique cognates could be viewed as a type of root cognates. However, we prefer to use 'root cognate' for those cases where the reflexes inherit the same form of the root and reserve 'oblique cognate' for the cases where this criterion is not necessarily met. Thus, strictly speaking we regard all cases of root cognates as also instances of oblique cognates, but practically speaking one would not typically call cases in which the same form of the root appears in two reflexes 'oblique cognates' because the more precise term 'root cognate' is available. As this example shows, oblique cognates are the result of extensive analogical and derivational developments; no single _état de langue_ is likely to have contained both _*pét-r-eh$_2$-_ and _*pt-er-ó-_ (Fellner & Hill 2019a, 168-169).

Oblique cognates arise primarily from non-concatenating morphology. The importance of accent and ablaut patterns to Indo-European morphology means that oblique cognates are very common in this family. The 15 etyma in Allen Nussbaum's (1986) account of words for 'head' and 'horn' in the older Indo-European languages all descend ultimately from _*ḱér-h$_2$/*ḱr-éh$_2$_, but none is entirely _lautgesetzlich_. The simplest case is Mycenaean Greek _kerā_ 'horn (material)' (< _*ḱér-eh$_2$_), either a reflex of the rectus stem with an analogical full-grade of the suffix or a reflex of the obliquus stem with an analogical full-grade of the root. In contrast, the pathway from the same proto-form to Latin _cerebrum_ (< _*ḱérh$_2$sro-_) requires six steps, which include a variety of morphological affixations, analogical derivational developments, and semantic changes. First, _*ḱér-h$_2$_, oblique _*ḱr-éh$_2$_ → _*ḱr-ḗh$_2$_, oblique _*ḱr̥-h$_2$-_ as a regular productive, so called ‘internal’, derivation (see discussion in Fellner & Hill 2019b, 117 n. 39 and cf., e.g., _*si̯éu̯H-mn̥_, oblique _*si̯uH-mén-_ in Skt. _syū́ma_ ‘band’ → _*si̯uH-mḗn_, oblique _*si̯uH-mn-_ in Gk. ὑμήν ‘membrane’) (Nussbaum 1986, 120, 134), accompanied by a change of meaning to ‘the head bone’. Second, the meaning shifted further to ‘skull, head’. Third, the analogy _*h$_2$eu̯s_ : oblique _*h$_2$us-es-_ ‘ear’ :: _*ḱr-ḗh$_2$_ : oblique X = _*ḱr̥h$_2$-es-_, led to the obliquus stem becoming _*ḱr̥h$_2$-es-_ (Nussbaum 1986, 214). Fourth, in Proto-Indo-European, in addition to the originally endingless locative stem (with its own ablaut grade different from the rectus and obliquus), there existed several affixal markers to characterize the locative, the most prominent being _*-i_ and _*-er_ (cf. Vedic _uṣás-i_ ‘at dawn’ (paradigmatic locative of _uṣas-_) < _*h$_2$us-és_ + _*-i_ next to (a substantive that arose by paradigmatic split of a locative) _uṣar-_ ‘thing at dawn’ < _*h$_2$us-s_ + _*-er_) the latter of which suffixed to our form gave _*ḱr̥h$_2$-s-er_ ‘on the head’ (Nussbaum 1986, 236). Fifth, this form was itself turned into an adjective with the adjective forming suffix _*-ó-_ to yield _*ḱr̥h$_2$-s-r-ó-_ ‘adj. in/at/on the head’ (cf. Vedic _usrá-_ ‘early’ < _*h$_2$us-s-r-ó-_ (Nussbaum 1986, 243)). In the final step, this adjective is nominalized with a change of accent to _*ḱérh$_2$sro-_ ‘thing on the head’ (Nussbaum 1986, 243); cf. Gk. λευκός ‘white’ : Gk. λεῦκος ‘white thing > whitefish’; Skt. _kṛśnás_ ‘black’ : Skt. _kṛ́śnas_ ‘black thing > black antelope’. Latin _cerebrum_ is the direct _lautgesetzliche_ outcome of _*ḱérh$_2$sro-_. The somewhat surprising change _*-sr-_ > _-br-_ is regular in Latin (see Weiss 2009, 163).

In Asian historical linguistics, many investigators reconstruct various alternate forms of the same root (see Blust 1990, 142-143 for Austronesian and Matisoff 1973, 123 for Trans-Himalayan). The pervasiveness of such reconstructed doublets itself suggests an inflectional morphological profile for the relevant proto-language (_pace_ LaPolla 2017, 40, 51).

4.4 Core cognate dimensions

The level of similarity between two cognates can be measured and visualised in three dimensions. A pair of comparanda may get a perfect score in phonological form on the y-axis, for instance, if all their phonological segments can be aligned and their sound correspondences are found in the correspondence set (C_t). However, one of the comparanda (or both of them), may have undergone various shifts in meaning and function, yielding a much lower similarity score on the x- and z-axes. The next section presents the diagnostics and proposed scoring metrics in detail.

5 Diagnosing cognates

In this section we propose a number of diagnostics to first of all determine whether comparanda are cognates and, second, if they are, what type and level of cognacy they represent. The first diagnostic test, described in section 5.1, is based on phonological form only. We next zoom in on the distinction between synonymous and non-synonymous cognates to establish the similarity of cognate pairs in terms of semantic similarity (section 5.2) as well as a number of morpho-syntactic and pragmatic variables (section 5.3).

5.1 Phonological alignment

Operationally the easiest metric to compute the level of cognacy is to focus on phonological similarity only. Naively, we could thus take the number of segments in word w_1 and word w_2 that are alignable (c_i) and found in the sound correspondence set of the two language comparanda (C_t) over the total number of alignable segments (*i*), i.e.

$$Cog(w_1, w_2) = \frac{\sum_{c_i}}{i}$$

This would work fine for examples like the ones shown in figure 1, where all comparanda have the same number of phonological segments and aligning the sound correspondences is straightforward. However, if we want to align Spanish *sol* with French *soleil* 'sun', the final segments of French *soleil* do not have any equivalent in Spanish *sol*. If we want to align these segments anyway, they will have zero as equivalents in Spanish. Since sound correspondences with zero are not found in the Spanish-French set of sound correspondences C_t, the result of the above equation would tell us Spanish *sol* and French *soleil* are only partially cognate. This in itself is not a bad result, but problems arise when the alignment of segments is less obvious.

As discussed in section 3 above, aligning segments of varying length is a non-trivial task to automate as in principle, without any prior knowledge, it is impossible to know where the zero segments should be added. A default 'end of word' approach would happen to work for *sol-soleil* but sometimes, zeros should be added to the beginning or right in the middle of words (e.g. in cases of epenthesis).

Ideally, we would calculate the number of (*unlautgesetzlich*) innovations that separate two forms, but this is only rigorously possible at an exceedingly advanced stage of research when C_t has been extended, tested and well-refined in terms of phonological conditioning. In the next sections we discuss how phonological alignment can be used as a diagnostic to determine the level of cognacy, ranging from strong ('strict cognates') to medium ('quasi-strict cognates') and weak cognates.

We propose instead to make the distinction between medium and weak cognates based on morphology, rather than phonology. In theory, we could propose a threshold of a minimum proportion of phonological segments that can be aligned as a cut-off point for medium cognates. However, this would make the distinctions more fluid and scalar making it harder to categorise comparanda. Therefore we propose a simple diagnostic for distinguishing medium cognates from weak cognates: if the comparanda are morphologically different and derived from different morphological proto-forms, they should be categorised as weak cognates. In the aforementioned comparanda Spanish *sol* (< *sōl*) vs. French *soleil* (< **sōliculus*, see REW, §8067), only the first part of the French word is etymologically derived from the same stem as the Spanish. The second part of the French *soleil* is derived from a Vulgar Latin diminutive *-iculus*, a morpheme which is not found in the history of Spanish *sol*. These can therefore not be medium cognates since not all morphemes are derived from the same proto-form; instead, they are weak cognates.

5.1.1 Strict cognates

Strict cognates are the strongest type of cognates, because all phonological segments of the cognate sets can be aligned, segment by segment, and the resulting correspondence patterns can be found in the permitted set of sound correspondences C_t.

If two forms descend from the same ancestor and have been perfectly transmitted in every segment from the proto-language, their pairwise segmental differences are explainable by regular sound change alone, and we can then arrange the words in a matrix where each word is placed in a row in such a way that regularly corresponding segments are placed in the same column (see figure 1 above), with segments not corresponding to any other segments (resulting from loss or epenthesis) being compared with null-segments (*gaps*), usually represented by a dash (−) symbol. Stated more formally:

- let w_1 be a word in language l_1 and w_2 a word in language l_2
- let c_i be {n, m} where n is the n^{th} segment of w_1 or a gap and m is the m^{th} segment of w_2 or a gap, and where $n \cup m \neq \emptyset$
- let C_t be the predefined set of all phonological correspondence patterns relating l_1 and l_2 that are deemed regular at time t
- if $\forall c_i \in C_t$, then w_1 and w_2 are strict cognates

Two strict (phonologically alignable) cognates might still not descend from the same inherited form; this is particularly a risk if the putative cognates are morphologically derived and the derivational morphology is itself cognate. To take an example, Brugmann (1881, 302) identifies Skt. *tyājáyāmi* ‘causes to quit, leave’ and Gk. σοβέω ‘scare away (birds), shoo (flies)’. The stems of both words reconstruct straightforwardly to **ti̯og*^{w}*-éi̯e-*. However, Watkins (1990, 297) suspects, presumably on the basis of its relatively late attestation and transparent semantics that *tyājáyāmi* ‘is productively formed and *pace* Brugmann does not make a true equation’ with σοβέω. In other words, Skt. *tyājáyāmi* provides evidence for the reconstruction of a root **ti̯og*^{w} and also provides evidence for the causative suffix **-éi̯e-*, but it does not directly support the reconstruction of a verbal stem **ti̯og*^{w}*-éi̯e-* in Proto-Indo-European. Rix (LIV, 643), by omitting *tyājáyāmi* from the descendants of **ti̯og*^{w}*-éi̯e-* concurs with Watkins. The lesson of this example is that the diagnostic criterion of alignability must be counterbalanced by the heuristics that late attestation and straightforward semantics (in morphologically derived words) weigh against a proposal of cognacy. Naturally, *tyājáyāmi* and σοβέω are still correctly regarded as cognates, but as partial cognates (section 4.3.1) rather than as strict cognates.

5.1.2 Labelling various medium cognates

Medium cognates are those cases where two words derived from the same proto-form have been affected by a non-phonological change, resulting in an exceptional status for an individual segment. In section 4.2 we listed two types of medium cognates: ‘quasi-strict’ (section 4.2.1) and ‘word equations’ (section 4.2.2). As discussed above, ‘quasi-strict cognates’ have three types of origins characterised by the manner in which one of their segments is affected by a non-phonological change, viz. through paradigm-internal analogy, contamination or interdialectal borrowing. Word equations form a somewhat separate category: these are cases derived from the same root continuing furthermore at least some inherited grammatical feature.

An etiological classification of the quasi-strict cognates could also be operationalised with the following diagnostics:

1. Check if the form participates in a paradigm that might provide a motivation for analogical change
2. Check if the form participates in a semantically coherent subsystem (e.g. numerals) of the type that is known to precipitate contamination
3. If neither of the first two check results in something promising, conclude it is likely to be a case of dialect borrowing

Ideally, one can provide independent evidence (e.g. facts about the historical phonology of the donor dialect) for the supposition of inter-dialectal borrowing, but this is often not possible and inter-dialectal borrowing can be seen as a 'catch-all' for the residue of as yet unexplained forms.

A word equation is to some extent discourse specific, since the two words compared must share an inherited category that the analyst is attempting to establish as present in the proto-language. As such, the computational identification of word equations is not necessarily sensible as a task.

5.1.3 Labelling 'partial' and 'oblique' weak cognates

As mentioned above (section 4.3), there are two types of weak cognates: partial and oblique cognates. Both can be distinguished from medium cognates because unlike medium cognates, not all morphemes are derived from the same morphological proto-form. Cases like Spanish *sol* vs Latin *soleil* above, where only one morpheme is derived from a different source are called 'Partial Cognates'. 'Oblique Cognates', on the other hand, are the result of extensive analogical and derivational developments. English *feather* (< **pétr-eh$_2$-*) and Greek πτερόν (< **pter-ó-*) are good examples of these as they exhibit different forms of the root as well as different suffixes (section 4.3.2). The most extreme form of 'oblique cognates' are examples like English *tooth* and Old Irish *dofúaid*, where only one phoneme in each of the forms can still be derived from the same root.

5.2 Semantic alignment

Once the level of cognacy is established based on the historical phonological similarity between the two comparanda, the next dimension of comparison is the semantics: if two forms are cognate, of whichever level, are they synonymous or non-synonymous? Since lexical semantics is inherently biased when comparing two words in different languages, the only way to automate this process objectively is through distributional semantics. In theory, this can be done through state-of-the-art NLP methods using diachronic word embeddings tracking the change of words in a particular language over time (see Hamilton, Leskovec & Jurafsky 2016, Kutuzov, Øvrelid, Szymanski & Velldal 2018, Bizzoni, Degaetano-Ortlieb, Menzel, Krielke & Teich 2019, Dubossarsky, Weinshall & Grossman 2017, Dubossarsky, Tsvetkov, Dyer & Grossman 2015, Dubossarsky, Weinshall & Grossman 2016). Since usual cross-linguistic methods are biased (as they rely on pre-established bilingual dictionaries), the only way to compare the semantic changes between the comparanda using diachronic word embeddings is by comparing the developments and rates of change in each of the languages. In practice, however, we face a number of difficulties working with scarcely attested stages of languages (cf. Meelen 2019, Fonteyn 2020, Felbur, Meelen & Vierthaler 2022). To get good results using diachronic word embeddings, we need large amounts of data at various stages/windows of the languages involved. When it comes to phonological reconstruction, we could therefore perhaps imagine comparing Modern Spanish *sol* to French *soleil*, vectorising stages of the languages all the way back to Classical Latin. Although this would require a large amount of preprocessing of the data in various stages (ensuring lemmatised and balanced, comparable corpora from which word embeddings are created), it is possible as long as there is enough data at each selected stage. When going beyond Latin, however, or when trying to reconstruct any proto-form, we have no data to work with, making comparison of diachronic word embeddings impossible. Further research in line with the work of Montariol & Allauzen (2019) on scarce data is necessary before these methods can be effectively extended to the work on historical reconstruction we are concerned with here.

When tracing the development of forms back to proto-languages, it is therefore better to rely on alternative methods for the time being. We propose that using colexification databases, such as CLICS (Rzymski & Tresoldi 2019) is currently the best way to diagnose the level of semantic similarity between cognates. The use of meaning 'concepts' is particularly useful. In its most simplified form, we only check whether cognates are listed as the same concepts and are thus synonymous (e.g. Dutch *stad* and German *Stadt* 'town'). Going one step further, we could diagnose different levels of 'non-synonymous' cognates, namely those that have undergone only slight changes in meaning for which a clear path of semantic change can be established and those that are completely different. The colexifica-

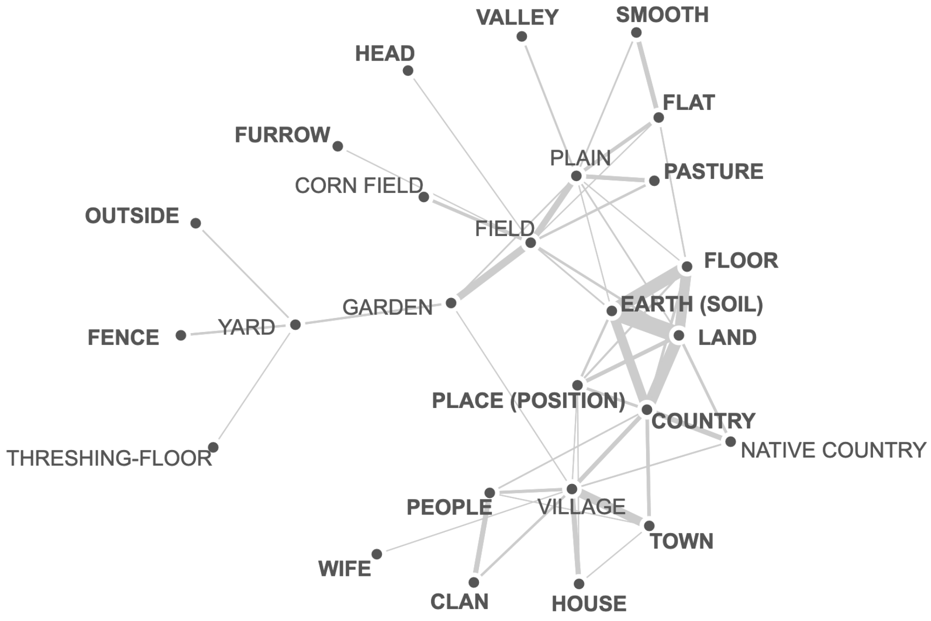

Figure 3: Subgraph from CLICS database, showing colexification strengths among the concepts 'town', 'fence', and 'garden'.

tions in the CLICS database can help with that, e.g. Dutch *land* 'country' and English *land*. When looking for the concept COUNTRY, the concept LAND is the first colexification with 217 links. Dutch *land* 'country' and English *land* are thus closely related even though they are not strictly synonymous. Dutch *tuin* '(fenced) garden', German *Zaun* 'fence' and English *town*, on the other hand, appear completely different at first sight. The concept TOWN, however, has a number of colexifications (e.g. VILLAGE, FORTRESS), and each of these colexifications can be linked by subgraphs, e.g. FENCE yielding the German *Zaun*. In turn, FENCE can be connected to YARD yielding the Dutch *tuin* 'garden' as a result (see figure 3 and table 1). These connections can thus be quantified depending on their colexifications, yielding options ranging from strict synonymous cognates to less strict (i.e. through one or more colexifications) and non-synonymous cognates.

As a concrete metric, we propose to count the number of edges that must be travelled to link the meaning of one cognate with the meaning of the other. A higher number of edges is a weaker semantic link. However, we want to count heavy-weighted edges for less, because they are widely attested co-lexifications. Consequently, we propose that we take the sum

Start node	End Node	Number of Colexifications
fence	yard	8
yard	garden	9
garden	village	3
village	town	66

Table 1: Colexification strengths along the path from 'fence' to 'town' in the CLICS database.

of edges, where each edge is counted as the inverse of its weight. Stated more formally:

- let w_1 be a word in language l_1 and w_2 a word in language l_2
- let $\mathfrak{w}_1$ and $\mathfrak{w}_2$ be the concepts in the Conception database that are mapped to as models of the denotation of w_1 and w_2
- let e_i be the i^{th} edge in the path that starts at $\mathfrak{w}_1$ and $\mathfrak{w}_2$
- let F_i be the weight[7] assigned to e_i in the CLICS database

We can then define the semantic closeness of w_1 and w_2 as follows:

$$Sem(w_1, w_2) = \sum_i \frac{1}{F_i}$$

To give a few examples:

$$Sem(land_{Dut.}, land_{Eng.}) = \frac{1}{127} = 0.0079$$

$$Sem(Zaun_{Ger.}, tuin_{Dut.}) = \frac{1}{8} + \frac{1}{9} = 0.2361$$

$$Sem(Zaun_{Ger.}, town_{Eng.}) = \frac{1}{8} + \frac{1}{9} + \frac{1}{3} + \frac{1}{66} = 0.5846$$

Note that *Sem(tooth*$_{Eng.}$*,dofúaid*$_{OIr.}$*)* cannot currently be calculated with this methodology because 'tooth' and 'eat' are not connected in the CLICS database. One can presume, however, that if the etymology is correct, then

[7] We use F_i, inspired by the F (force) of physics, in order to avoid the confusion of using w, which is already used for 'word'.

a future edition of the database will link these two graphs, albeit very weakly, i.e. *Sem(tooth*$_{Eng.}$*,dofúaid*$_{OIr.}$*)* will be a large number.

This methodology of course relies on the correct mapping of words to their closest meanings in the Concepticon Database. The Dutch word *tuin* includes the connotation that the garden is enclosed with a fence, thus it is actually semantically closer to German *Zaun* then the mapping to GARDEN in the CLICS database makes clear.

5.3 Syntactic and pragmatic alignment

Apart from similarity in phonological form and meaning, cognates can also be more or less similar in syntactic and pragmatic function. Reanalyses and grammaticalisation processes as well as other syntactic and pragmatic developments between the proto-language and the present-day form can change the function of cognates, just as much as its phonological form or meaning can change over time. In principle, the similarity of any number of functional parameters can be measured. For the present paper, we focus on the main morpho-syntactic categories as well as their subtypes.

First we determine the core part of speech for each cognate, i.e. its prevalent morpho-syntactic function in the language at its current state. Although nouns and verbal stems are most commonly compared, in principle many core parts of speech can occur, as in the following (non-exhaustive) list:

- verbs (verbal roots or stems)
- nouns
- pronouns
- numerals
- adverbs
- adjectives
- adpositions (prepositions, postpositions)
- determiners (articles, demonstratives)
- particles (negation, focus, question, etc)

Although there are exceptions, especially with weak cognates (e.g. English *tooth* vs Old Irish *dofúaid*), often the core parts of speech of each of the cognates will be the same. In the following sections we therefore present a number of diagnostics to distinguish between various subtypes, which can reveal more detailed changes of syntactic or pragmatic functions. As long as all of the cognates under comparison are submitted to the same diagnostic tests, when testing similarity of function we can go into any level of detail. In practice a high level of detail may only be useful in automated procedures where the similarity of large amounts of cognates (in form, meaning and function) is computed. When manually comparing cognates a more superficial level of detail, e.g. a simple comparison on the part-of-speech level could be sufficient information to determine whether cognates are similar or not in terms of their function. The following subsections provide some examples of how to classify some parts of speech further based on their core functions.

5.3.1 Verbs

Verbs can be classified as intransitives, (optionally) transitives or ditransitives depending on the number of arguments (one, two or three respectively) they take. Intransitive verbs can furthermore be split into unergatives and unaccusatives depending on the nature of their one core argument. Diagnostics for distinguishing between these subcategories can vary from language to language. In section 2.3 we presented detailed examples from Dutch and English, but these, to a certain extent, can be applied to other languages as well, e.g. German or Italian and French.

5.3.2 Nouns

Nouns could be divided into various subtypes as well, but for present purposes, we limit ourselves to a basic distinction between mass and count nouns. Diagnosing count nouns can be easily done by testing whether plural markers (affixes, determiners, etc) and numeral modifiers of two and three or higher are allowed. In English, the count noun *cloth*, can be distinguished from *clothing*, because *three cloth**s*** is possible, whereas **three clothing**s*** is not. Note that certain mass and collective nouns in many languages can be unitised, however, when plural interpretations are derived from the unit it can be measured in. Examples of these in English are *rice* or *milk*, where *two rices/milks* in fact denotes ‘two bowls of rice’ and ‘two glasses of milk’ respectively.

5.3.3 Adverbs

Adverbs that are adjuncts (e.g. adverbs or time or place) often exhibit a certain amount of distributional freedom (cf. Bonami, Godard & Kampers-

Manhe 2004); the function and distribution of scopal adverbs, on the other hand, is more restricted. Various functions of adverbs could be tested for in theory, but we limit ourselves to one core example known from traditional classification of adverbs in the literature (e.g. Jackendoff 1972), i.e. their scope. Many adverbs in English and other languages have either broad or narrow scope. Some, however, can have both narrow scope (i.e. just over the verb phrase: 'VP scope') and broad scope (i.e. over the entire proposition: 'CP scope' or so-called 'S adverbs'). Examples of each of these types are given in (3), whereas example (4) shows certain adverbs, like English *hopefully*, could have either function:

(3) a. He *completely* ate the cheese. [VP adverb]
b. He *evidently* ate the cheese. [S adverb]

(4) a. He *hopefully* walked home, thinking this time he finally made a difference. [VP adverb]
b. *Hopefully*, the weather will be nice tomorrow. [S adverb]

The syntactic position of these adverbs that can function as either VP or S adverbs determines their scope. The VP adverb *hopefully* in example (4-a), which is modifying the verb only, cannot occur sentence-initially. If it does, as shown in (4-b), its scope widens to modify the entire proposition. In the same vein, Potsdam (n.d.), for example, gives the following examples showing broad-scoped adverbs must precede narrow-scoped adverbs in English:

(5) a. Hulk Hogan [evidently]$_S$ [completely]$_{VP}$ annihilated his opponent.
b. *Hulk Hogan [completely]$_{VP}$ [evidently]$_S$ annihilated his opponent.

In addition to these VP and S adverbs, Jackendoff (1972) identifies a third type, which have the positional distribution of neither of the former two classes. Potsdam (n.d.) labels these 'E(xtent) Adverbs' because they describe the extent to which a situation holds. Examples of these in English are *merely, hardly, scarcely* etc. More detailed distinctions between different types of adverbs cross-linguistically, depending on their positional distribution are made by, among others, Cinque (1999) and Rizzi (2004).

5.3.4 Other particles

'Other particles' come in various shapes and forms and are deliberately not specified here further to facilitate cross-linguistic comparison. Depending on the language, any ‘markers’, ‘operators’, ‘particles’ or any remaining parts of speech can convey pragmatic functions. We briefly discussinformation-structural and speech-act features here.

There are three core dimensions of information structure:

- focus vs background
- topic vs comment
- given vs new Information

These features can be expressed in the language through phonology (e.g. intonational phrases indicate certain types of topics in English, Japanese and German, cf. Krifka & Musan 2012, 34), morphology (e.g. suffixes to mark VP focus such as *-go* in Chadic, cf. Hartmann & Zimmermann 2007), syntax (e.g. various V2 and cleft orders in Middle Welsh, cf. Meelen 2016, chapter 5) and lexical items and particles. In this section, we focus on lexical and functional items as these are most likely to be reconstructed, however, establishing the cognacy of morphological affixes is also possible.

In Dutch, for instance, *ook* ‘also’ often functions as a focus marker. Old Frisian *āk* ‘also, even’ and German *auch* are both adverbs that function as a focus markers too. These are strict cognates as they exhibit perfect phonological alignment. They are furthermore synonymous, since their semantic and pragmatic functions are the same as well. Old English *ēac* ‘with, besides’, however, is a preposition. It thus scores slightly lower on the functional similarity scale.

A good example of a functional marker exhibiting speech-act features are pragmaticalised Norwegian *sánn*, German *son* and Dutch *zo’n* ‘such an X’ (of the kind that we both know). Speech-act features can be oriented towards the speaker, hearer or both participants. Kinn & Meelen (forthcoming) argue that in both Norse and Dutch, the new pragmatic function relates Norwegian *sånn* and Dutch *zo’n* to both Speaker and Hearer features, yielding its new pragmatic ‘recognitional’ interpretation to mark that the noun phrase it modifies is in the common ground of both. Originally, both items are derived from demonstratives with a deictic function. In this case then German *son* and Dutch *zo’n* are not just strict cognates exhibiting perfect phonological alignment, they are also synonymous in terms of semantics and pragmatics. Other examples of cognates whose pragmatic functions can be meaningfully compared in this way are, for example, a number of evidentiality markers that exhibit speech-act features as well.

5.4 Metrics for automatic cognate detection

Figure 4 presents a simplified representation of the core variables: of form, meaning and function. Note that this 3-dimensional visualisation is just a simplification. In practice, each subvariable could be a metric and each of the variables could get a weight to give prominence to whichever factors are deemed more important in the comparison, e.g. phonology and semantics. Synonymous cognates differ from non-synonymous cognates because they exhibit similarity in meaning and function. In the above sections on semantic, morpho-syntactic and pragmatic similarity, we gave a number of diagnostics to test whether cognates exhibit functional similarity or not. Each of these could be linked to a clear scoring metric to facilitate automatic cognate similarity comparison. Semantic similarity can be measured by using the CLICS database of concepts, checking how distant colexifications are. When it comes to morpho-syntactic and pragmatic variables, similarity scores can be established for each of the features under comparison, e.g. cognates that are both verbal stems, but differ in level of transitivity, are more similar than cognate pairs consisting of nouns and verbs, but less similar than those pairs that are both unergative intransitive verbs. A scoring scale can thus be established for each cognate language pair under investigation.

Figure 4 shows the three core dimensions and two samples of resulting three-dimensional planes: the smaller the surface of the plane (i.e. the closer all three scores are to (0,0,0) or an arbitrary number, e.g. 100%, the stricter (qua phonological form) and more synonymous (in meaning and function) the cognates are.

6 Conclusion

We hope that the foregoing discussion has succeeded in tightening up what is meant by ‘cognate’ in historical linguistics and partially formalizing the evaluation of whether two forms are cognate. This (partial) formalization of cognacy and its associated workflow serves as one small subcomponent of an overall formalization of the comparative method. The need for such an overall formalization is now widely recognized, both for its inherent intellectual merits and because a computational implementation of the comparative method is the only way that we can resonably hope for non-Indo-European language families to become as well understood as Indo-European.

Historical linguists typically speak as if cognancy is a binary relationship. This conceit is perhaps a convenient simplification in the context of

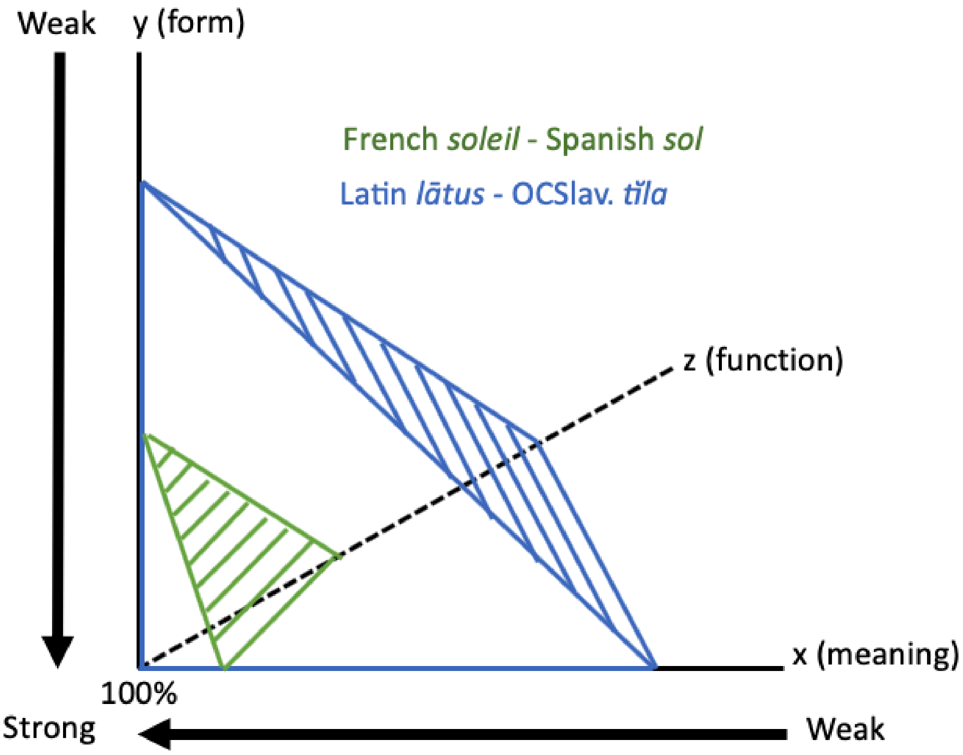
Weak
y (form)
French *soleil* - Spanish *sol*
Latin *lātus* - OCSlav. *tĭla*
z (function)
x (meaning)
100%
Strong
Weak

Figure 4: Cognacy variables in 3 dimensions indicating similarity in form, meaning and function.

traditional historical linguistics, but it is a dangerous misunderstanding when taken for granted in machine readable datasets. In particular, future phylogenetic work would merit from deploying a more sophisticated model of cognacy.

Comments invited

PiHPh relies on post-publication review of the papers that it publishes. If you have any comments on this piece, please add them to its comments site. You are encouraged to consult this site after reading the paper, as there may be comments from other readers there, and replies from the author. This paper's site is here:

http://dx.doi.org/10.2218/pihph.7.2022.7405

Acknowledgements

The authors would like to thank Johann Mattis-List for valuable comments on an earlier version of this paper. We also acknowledge the ERC grant 'ASIA: Beyond Boundaries: Religion, Region, Language and the State' (2014-2020, ID: 609823), under the auspices of which Marieke Meelen and Nathan W. Hill first began working on this paper. Research for this paper was also supported by the Austrian Science Fund (FWF): project number Y-1044.

Author contact details

Marieke Meelen
University of Cambridge
Trinity Hall
Trinity Lane
Cambridge CB2 1TJ
United Kingdom

mm986@cam.ac.uk

Nathan W. Hill
Trinity College Dublin
Trinity Centre for Asian Studies
Dublin 2
Ireland

nathan.hill@tcd.ie

Hannes Fellner
University of Vienna
Department of Linguistics
Sensengasse 3a
1090 Vienna
Austria

hannes.fellner@univie.ac.at

References

Alexiadou, Artemis, Elena Anagnostopoulou & Martin Everaert. 2004. *The unaccusativity puzzle: explorations of the syntax-lexicon interface*.

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

Anttila, Raimo. 1989. *Historical and comparative linguistics*, vol. 6. John Benjamins Publishing.

Bizzoni, Yuri, Stefania Degaetano-Ortlieb, Katrin Menzel, Pauline Krielke & Elke Teich. 2019. Grammar and meaning: analysing the topology of diachronic word embeddings. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 175–185.

Blust, Robert. 1990. Patterns of sound change in the Austronesian languages. In Philip Baldi (ed.), *Linguistic change and reconstruction methodology*, 231–270. Berlin; New York: Mouton de Gruyter.

Bonami, Olivier, Danièle Godard & Brigitte Kampers-Manhe. 2004. Adverb classification. *Handbook of French semantics*. 143–184.

Brugmann, Karl. 1881. Griechische etymologien. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen* (25). 298–307.

Bynon, Theodora. 1977. *Historical Linguistics*. Cambridge: Cambridge University Press.

Campbell, Lyle. 2004. *Historical linguistics: an introduction*. 2nd edn. Edinburgh: Edinburgh University Press.

Chang, Will, Chundra Cathcart, David Hall & Andrew Garret. 2015. Ancestry-constrained phylogenetic analysis ssupport the Indo-European steppe hypothesis. *Language* 91(1). 194–244.

Cheng, Lisa Lai-Shen & Norbert Corver. 2013. *Diagnosing syntax*, vol. 46. Oxford University Press.

Cinque, Guglielmo. 1999. *Adverbs and functional heads: a cross-linguistic perspective*. Oxford University Press on Demand.

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

Crowley, Terry. 1982. *The Paamese language of Vanuatu*. Canberra, A.C.T., Australia: Dept. of Linguistics, Research School of Pacific Studies, Australian National University.

Crowley, Terry & Claire Bowern. 2010. *An introduction to historical linguistics*. 4th edn. Oxford: Oxford University Press.

Crystal, David. 2008. A dictionary of linguistics and phonetics. malden. *MA: Blackwell*.

Davies, Anna Morpurgo. 1998. *Nineteenth-century linguistics*. Giulio Lepschy (ed.) (History of Linguistics IV). London: Longman.

Dimmendaal, Gerrit J. 2011. *Historical linguistics and the comparative study of African languages*. Amsterdam: John Benjamins Publishing Company.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer & Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Networds*, 66–70.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2016. *Verbs change more than nouns: A bottom-up computational approach to semantic change*. Unpublished Manuscript. https://www.academia.edu/25793914.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1136–1145.

Felbur, Rafal, Marieke Meelen & Paul Vierthaler. 2022. Crosslinguistic semantic textual similarity of Buddhist Chinese and Classical Tibetan. *Journal of Open Humanities Data* 8(1). 23, 1–14.

Fellner, Hannes & Nathan W. Hill. 2019a. The differing status of reconstruction in Trans-Himalayan and Indo-European. *Cahiers de Linguistique – Asie Orientale* 48(2). 159–172. https://brill.com/view/journals/clao/48/2/article-p159_5.xml.

Fellner, Hannes & Nathan W. Hill. 2019b. Word families, allofams, and the comparative method. *Cahiers de linguistique – Asie Orientale* 48(2). 91–124.

Fonteyn, Lauren. 2020. What about grammar?: Using BERT embeddings to explore functional-semantic shifts of semi-lexical and grammatical constructions. In *Proceedings of CHR 2020: Workshop on computational humanities research*, 257–268. CEUR-WS.

Fortson, Benjamin W. 2010. *Indo-Eeuropean languages and culture: an introduction*. Malden, Oxford & Victoria: Blackwell.

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

Gray, Russell D. & F. M. Jordan. 2000. Language trees support the express-train sequences of Austronesian expansion. *Nature* (405). 1052–1055.

Hale, Mark. 2007. *Historical Linguistics: Theory and Method*. 1st edn. (Blackwell Textbooks in Linguistics). Malden, Oxford & Victoria: Wiley-Blackwell.

Hamilton, William L, Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Hartmann, Katharina & Malte Zimmermann. 2007. Focus strategies in Chadic–the case of Tangale revisited. *Studia Linguistica* 61(2). 95–129.

Hill, Nathan W. 2014. Grammatically conditioned sound change. *Language and Linguistics Compass* 8(6). 211–229. http://eprints.soas.ac.uk/18595/.

Hill, Nathan W. 2019. *The historical phonology of Tibetan, Burmese, and Chinese*. London: Cambridge University Press.

Hill, Nathan W. & Johann-Mattis List. 2017. Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting* 3(1). 47–76.

Hock, Hans Henrich. 1991. *Principles of historical linguistics*. Berlin: Mouton de Gruyter.

Hockett, Charles F. 1967. Where the tongue slips, there slip I. In *To honor Roman Jakobson*, 910–936. Mouton.

Hoenigswald, Henry Max. 1960. *Language change and linguistic reconstruction*. 4. Aufl. 1966. Chicago: The University of Chicago Press & Univ. of Chicago Press.

Huehnergard, John. 2004. Afro-Asiatic and Semitic languages. In Roger D. Woodard (ed.), *The cambridge encyclopedia of the world's ancient languages*, 225–246. Cambridge: Cambridge University Press.

Jackendoff, Ray S. 1972. Semantic interpretation in generative grammar.

Jasanoff, Jay H. 2003. *Hittite and the Indo-European verb*. Oxford: Oxford University Press.

Kinn, Kari & Marieke Meelen. Forthcoming. Formalising pragmaticalisation in Dutch and Norwegian DPs.

Koch, Harald. 1996. Reconstruction in morphology. In Mark Durie (ed.), 218–263. New York: Oxford University Press.

Koch, Harold & Luise Hercus. 2013. Obscure vs. transparent cognates in linguistic reconstruction. In Robert Mailhammer (ed.), *Lexical and structural etymology*, 33–51. Berlin & New York: de Gruyter.

Krifka, Manfred & Renate Musan. 2012. Information structure: overview and linguistic issues. *The expression of information structure* 5. 1–44.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Labat, Sofie & Els Lefever. 2019. A classification-based approach to cognate detection combining orthographic and semantic similarity information. In *Recent advances in natural language processing 2019*, 603–611.

LaPolla, Randy. 2017. Overview of Sino-Tibetan morphosyntax. In Randy J. Lapolla & Graham Thurgood (eds.), *The sino-tibetan languages*, 40–69. Routledge.

List, Johann-Mattis. 2012. Multiple sequence alignment in historical linguistics: A sound class based approach. In Enrico Boone, Kathrin Linke & Maartje Schulpen (eds.), *Proceedings of ConSOLE XIX*, 241–260.

List, Johann-Mattis. 2014. *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

List, Johann-Mattis. 2019. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* 45(1). 137–161. https://www.aclweb.org/anthology/J19-1004.

List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi & Robert Forkel. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution* 3(2). 130–144.

Matisoff, J. A. 1973. Tonogenesis in Southeast Asia. In Larry H. Hyman (ed.), *Consonant Types and Tone*, 71–95. Los Angeles: UCLA.

Meelen, Marieke. 2016. *Why Jesus and Job spoke bad Welsh: The origin and distribution of V2 orders in Middle Welsh*. Utrecht: LOT dissertation series.

Meelen, Marieke. 2019. Darling, dukeling, duckling: How historical corpora can verify predicted pathways of language change. Keynote talk at the Cambridge Language Sciences Symposium, 19 November 2019.

Meyer-Lübke, Wilhelm (comp.). 1911. *Romanisches etymologisches Wörterbuch* (Sammlung romanischer Elementar- und Handbücher 3.3). Heidelberg: Winter.

Montariol, Syrielle & Alexandre Allauzen. 2019. Empirical study of diachronic word embeddings for scarce data. *arXiv preprint arXiv:1909.01863*.

Nussbaum, Alan J. 1986. *Head and horn in Indo-European. The words for "horn," "head," and "hornet"*. Berlin & New York: de Gruyter.

Potsdam, Eric. N.d. A syntax for adverbs. In *The Proceedings of the 1998 Western Conference on Linguistics (WECOL98)*.

Rama, Taraka & Johann-Mattis List. 2019. An automated framework for fast cognate detection and bayesian phylogenetic inference in computational historical linguistics. In *57th Annual Meeting of the Association for Computational Linguistics*, 6225–6235. Association for Computational Linguistics.

Ringe, Donald. 2006. *A linguistic history of English*. Vol. 1: *From Proto-Indo-European to Proto-Germanic*. Oxford: Oxford University Press.

Rix, Helmut (ed.). 2001. *LIV. Lexikon der Indogermanischen Verben: Die Wurzeln und ihre Primärstammbildungen*. In collab. with Martin Kümmel, Thomas Zehnder, Reiner Lipp & Brigitte Schirmer. Wiesbaden: Reichert.

Rizzi, Luigi. 2004. *The structure of CP and IP: The cartography of syntactic structures volume 2: The cartography of syntactic structures*, vol. 2. Oxford University Press.

Rzymski, Christoph & Tiago Tresoldi. 2019. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. https://clics.clld.org/.

Schwink, Frederick. 1994. *Linguistic typology, universality and the realism of reconstruction*. Washington: Institute for the Study of Man.

Thurneysen, Rudolf. 1946. *A grammar of Old Irish*. Trans. by D. A. Binchy & Osborn Bergin. Dublin: School of Celtic Studies, Dublin Institute for Advanced Studies.

Trask, Robert Larry (comp.). 2000. *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh University Press.

Trask, Robert Larry. 2015. *Trask's historical linguistics*. Robert McColl Millar (ed.). 3rd edn. London & New York: Routledge.

Vine, Brent. 1993. Greek -ίσκω and indo-european "*-isk̂e/o-". *Historische Sprachforschung / Historical Linguistics* 106(1). 49–60. http://www.jstor.org/stable/40849080.

von Mengden, Ferdinand. 2008. Paul Georg Meyer, *Synchronic English Linguistics: An Introduction*. *Anglia-Zeitschrift für englische Philologie* 126(1). 114–118.

Watkins, Calvert. 1990. Etymologies, equations, and comparanda: Types and values, and criteria for judgment. In Philip Baldi (ed.), *Linguistic change and reconstruction methodology*, 289–303. Berlin; New York: Mouton de Gruyter.

Weiss, Michael. 2009. *Outline of the historical and comparative grammar of Latin*. Ann Arbor: Beech Stave Press.

Weiss, Michael. 2014. The comparative method. In Claire Bowern & Nicholas Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 1st edn. (Routledge Handbooks in Linguistics), 127–145. New York: Routledge.