[Draft chapter, please cite as: List, Johann-Mattis (under review): Automatic methods for the investigation of language contact situations. Draft, submitted to Hickey, Raymond (ed): Handbook of Language Contact, Wiley-Blackwell, 2nd edition. 1-17]

Automatic methods for the investigation of language contact situations

0 Abstract

While language contact has so far been predominantly studied on the basis of detailed case studies, the emergence of methods for phylogenetic reconstruction and automatic word comparison -- as a result of the recent quantitative turn in historical linguistics -- has also resulted in new proposals to study language contact situations by means of automatic approaches. The chapter will provide a concise introduction to the most important approaches which have been proposed in the past, focusing on approaches that use (A) phylogenetic networks to detect reticulation events during language history, (B) sequence comparison methods in order to identify borrowings in multilingual datasets, and (C) arguments for the borrowability of shared traits to decide if traits have been borrowed or inherited.

1 Introduction

The past two decades have seen a drastic increase of quantitative applications in historical linguistics and linguistic typology. On the one hand, this *quantitative turn* (Geisler and List 2013) is reflected in multiple articles dealing with the automation of formerly exclusively manual tasks, including phylogenetic reconstruction and linguistic dating (Gray and Atkinson 2003, Holman et al. 2011, Chang et al. 2015), word comparison (Kondrak 2000, Prokić et al. 2009, List 2014, List et al. 2018), polysemy and semantic change (Steiner et al. 2011, List et al. 2013, Dellert 2016, Eger and Mehler 2016), and regular sound correspondences (Kondrak 2002, Brown et al. 2013, List 2018). The application of these methods was largely favored by the compilation of large cross-linguistic databases, offering large-scale accounts on typological structures (Polyakov and Solovyev 2006, Dryer et al. 2013), lexical cognates (Starostin 2008, Greenhill et al. 2008, Matisoff 2015), lexical data in general (Key and Comrie 2016, Dellert and Jäger 2017, Kaiping and Klamer 2018), phoneme inventories (Maddiesen et al. 2013, Moran et al. 2014), and polysemies (List et al. 2014, List et al. 2018).

Given the importance of language contact for the study of language history and linguistic typology, it is not surprising that automatic approaches to study language contact were also proposed in the past. In contrast to the numerous studies dealing with language history or universals, however, the majority of studies dealing with language contact scenarios has been restricted to the use of Neighbor-Nets (splits networks, Bryant and Moulton), as implemented by the SplitsTree software package (Huson 1998), which can be conveniently used with various types of linguistic data, including lexical data (Bryant et al. 2005, Hamed and Wang 2006, Bowern 2010), phonetic data (McMahon et al. 2007, Heggarty 2010, Prokić 2010), and typological data (Daval-Markussen and Bakker 2011, Széto et al. 2018).

While splits networks are a useful way to provide a summary of a dataset for the purpose of *exploratory data analysis* (Morrison 2014), they are of limited use in studying language contact phenomena directly, although scholars often state this otherwise in the literature. The reason for the limited use of splits networks in the study of language contact or language history is that they do not allow to infer *which traits* have been shared or influenced as a result of contact. For this reason, this chapter won't deal with splits networks and similar approaches to exploratory data analysis. Instead, it will focus on those methods which allow for a more concrete interpretation of the findings, be it by yielding explicit historical scenarios that allow for horizontal transmission, or by identifying explicit instances of borrowing among unrelated languages.

In the following, we will first look at the general problem of distinguishing contact-induced similarities from similarities resulting from universal tendencies of language change or genetic relatedness. We will then discuss how these problems are dealt with in classical, non-automatic frameworks. The classical approaches will then be contrasted by the most promising automatic techniques that have been proposed so far. Among these are (A) enhanced methods for phylogenetic reconstruction using phylogenetic networks, (B) sequence comparison techniques applied to multilingual wordlists, and (C) borrowability arguments that help to distinguish borrowed from inherited traits. We will conclude by looking at future chances and challenges for the application of computational approaches to study language contact situations.

2 Similarities and Language Contact

No matter whether one is interested in inherited or borrowed traits among languages: without resorting to some notion of *similarity of traits* across languages, it is not possible to study historical language relations. Depending on what traits (or what *comparative concepts* in the sense of Haspelmath 2010) we inspect, languages can be strikingly similar in various ways. They can share similar words, but also similar *structures*, be they lexical or grammatical. While some similarities may give us concrete hints regarding the shared history between two or more languages, many of the similarities we can observe among languages are coincidental or based on general ("universal") tendencies observed in the languages of the world. More systematically, we can distinguish similarities that are

1. coincidental (simply due to chance),
2. natural (being grounded in human cognition),
3. genealogical (due to common inheritance), and
4. contact-induced (due to lateral transfer).

As an example for the first type of similarity, consider Modern Greek θεός [θɛɔs] ‘god’ and Spanish *dios* [diɔs] ‘god’. Both words look similar and sound similar, but this is a sheer coincidence. This becomes clear when comparing the oldest ancestor forms of the words which are reflected in written sources, namely Old Latin *deivos*, and Mycenaean Greek *thehós* (Meier-Brügger 2002: 57f). As an example for the second type, consider Chinese *māmā* 媽媽 'mother' vs. German *Mama* 'mother': both words are strikingly similar, but not because they are related, but because they reflect general principles of early language acquisition, which usually starts with vowels like [a] and the nasal consonant [m]

(Jakobson 1960). An example for genealogical similarity are German *Zahn* and English *tooth*, both going back to a Proto-Germanic form *tanθ-. Contact-induced similarity is reflected in English *mountain* and French *montagne*, since the former was borrowed from the latter.

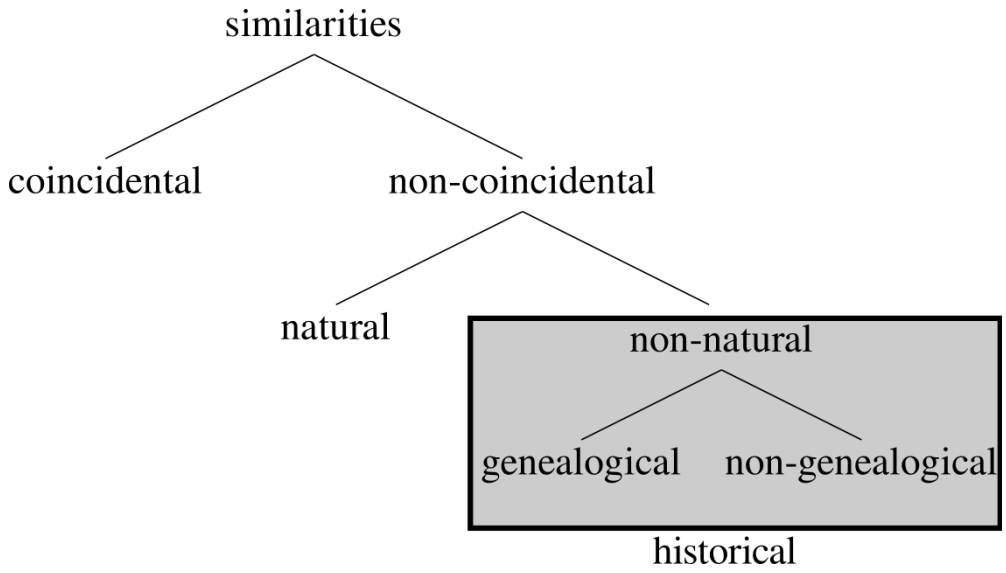

Figure 1: Similarities in linguistics.

Following List (2014: 56), we can display these similarities in a decision tree, as provided in Figure 1. In this figure, the last two types of similarity are highlighted in order to indicate that they are *historical*, reflecting individual scenarios of language change. Natural and coincidental similarities, on the other hand, are not indicative of language history.

When trying to infer similarities resulting from language contact, it is crucial to distinguish between the four different types of similarities. At times this can be rather trivial, especially if we can rule out from the beginning that languages sharing similar traits are genetically unrelated. In such cases, all that needs to be shown is that that similarities are unlikely to have evolved independently. If two or more genetically unreleated languages Identifying the dynamics of language contact among genetically closely related languages, on the other hand, is much more difficult.

3 Classical approaches to the study of contact situtations

The most straightforward way to study language contact is by means of *direct evidence*, based on direct observation, or the comparison of different stages of a language as they are reflected in written sources. The fact that Guǎngzhōu Chinese [tʰai³³ iœŋ²¹] 太陽 "sun" is a recent borrowing from Mandarin Chinese, for example, is easy to prove when comparing modern sources of the dialect with older ones. Thus, while the *Chinese dialect vocabularies* (Běijīng University 1964), a collection of dialect readings from the 1960s, lists only the form [jit²²tʰɐu²¹₃₅] 熱頭, a common innovation in the Yuè dialects, while more recent vocabulary collections, such as Liú et al. (2007) list exclusively the former form. Thus, if languages are well-documted in written sources, we can often directly see when a word enters

their lexicon.

If there is no direct evidence for linguistic interference, scholars need to resort to indirect techniques to prove that a certain trait in a given language arose from contact. In contrast to general language change, which proceeds in a largely regular manner, contact-induced change processes can be seen as disruptive and chaotic events that may occur but might as well not occur in the history of a language. As a result, the strategies that scholars use to detect and prove scenarios of lexical and grammatical interference build on the identification of historical similarities that are in conflict with a genealogical explanation. This requires, of course, to demonstrate first that the similarities under question are indeed historical (as opposed to natural or coincidental similarities as discussed above).

While the discipline of historical linguistics has developed sophisticated techniques to prove that language similarities are genealogical, the techniques for identifying contact-induced similarities are less homogenous and usually involve the detailed sifting of multiple pieces of evidence that achieve their convincing power only in combination. In this regard, techniques for the identification of linguistic inference are not much different from other, more specific, types of linguistics reconstruction, such as the reconstruction of the pronunciation of ancient languages based on "philological reconstruction" (Jarceva 1990, see Sturtevant 1920 for the pronunciation of Greek and Latin, or Baxter and Sagart 2014 for the reconstruction of Old Chinese phonology), the reconstruction of detailed etymologies (Malkiel 1954), or the reconstruction of syntax (see Willis 2011).

By nature, "cumulative-evidence arguments" (Berg 1998), or arguments based on *consilience* (Whewell 1840, Wilson 1998) are much more difficult to formalize than clear-cut procedures and unified tests that yield simple, binary results. As a result, it is often very difficult to formalize what scholars do in concrete in order to come to their conclusions. Despite these difficulties in determining exact workflows to deal with linguistic interference, we can identify a couple of *proxies* that scholars use to assess whether a given trait shared between two or more languages has borrowed or not.

One important class of hints are *conflicts* with genealogical explanations. A first type of conflicts is represented by similarities shared among unrelated or distantly related languages. Given that the English word *mountain* is reflected only in English, with similar words in the Romance languages, denoting the same concept, we could take this alone as evidence that the English word was borrowed. Since these conflicts here arise from the supposed phylogeny of the languages under consideration, we can speak of *phylogeny-related arguments* for interference.

A second conflict involves the traits themselves, most prominently observerd in the case of irregular sound correspondence patterns. German *Damm*, for example, is related to English *damb*, but since the expected correspondence for cognates between English and German would yield a German reflex *Tamm* (as it is still reflected in Old High German, see Kluge 2002), we can take this as evidence that the modern German term was borrowed (Pfeifer 1993). We can call these cases *trait-related arguments* for contact.

A third type of argument for interference can be derived from distributional properties of the shared traits. For example, if we observe that one language shares many words with another language, but all

words belong to a similar semantic field, such as, for example, religious or technical terms, this is usually also seen as a strong indicator of borrowing, since we would expect that related languages share words across many different semantic fields, specifically pertaining to the realm of *basic vocabulary*. We can call these *distribution-based arguments* for contact.

Note that these arguments for interference based on different types of conflict can be used for structural traits (such as grammar or phonology) as well. The lack of an infinitive in Balkan languages, such as Bulgarian and Greek (Friedman 2007: 208), for example, reflects a phylogeny-related conflict. The irregular plural ending *-a* in German *Lexikon* (plural *Lexika*), reflects (among others) a trait-related conflict, but it is also distribution-related, given its extremely limited scope. In general, however, it seems that phylogeny-related arguments prevail as a type of evidence for structural interference.

In addition to the observation of conflicts within shared traits, two further types of evidence are of great importance for the inference of language contact situations in the absence of direct evidence. The first one is areal proximity, and the second one is the assumed *borrowability* of traits. Given that language contact requires the direct contact of speakers of different languages, it is self-evident that areal proximity, including proximity by means of the travel of speakers via trade relations, is a necessary argument when proposing contact relations between different language varieties.

Given that direct evidence confirms that linguistic interference does not act to the same degree on all levels of linguistic organisation, the notion of *borrowability* also plays an important role, although scholars tend to have different opinions about borrowability in concrete. Most scholars would probably agree with the borrowability scale proposed by Aikhenvald (2007: 5), which ranges from "inflectional morphology" and "core vocabulary", representing aspects that are relatively resistent to borrowing, up to "discourse structure" and the "structure of idioms", representing aspects that are relatively easy to borrow. How core vocabulary can be defined, and how the borrowability of individual lexical comparative concepts can be determined and ranked, however, has been subject to controverse discussions (Starostin 1995, Lee and Sagart 2008, Tadmor 2008, Zenner et al. 2014).

4 Computational Approaches to Study Language Contact

Despite the large number of quantitative applications in historical linguistics during the last two decades, computational approaches to infer contact situations are still in their infancy. As of now, none of the few approaches that have been proposed in the past can compete with the classical methods based on the manual investigation of shared similarities by experienced experts. The reasons for the lack of approaches are twofold. First, given the multiple types of evidence employed by the classical approaches, the formalizion of the problem of borrowing detection is difficult. Second, given the limited number and suitability of datasets annotated for different types of linguistic interference (the World Loanword Database by Haspelmath and Tadmor 2009, for example, contains only 40 language varieties, and none of them is accompanied by phonetic transcriptions), scholars have a hard time in developing algorithms, since they lack data for testing and training.

In principle, all algorithms for the inference of language contact situations that have been proposed so far, make use of the same strategies that are also used in the classical approaches. Thus, they first infer

or determine shared traits among two or more languages, and then determine conflicts in these traits, taking areal closeness and borrowability into account. In contrast to classical approaches, which combine different types of evidence, computational approaches are usually restricted to one type.

The automatic methods proposed in the past can be divided into three different classes. The first class employs phylogeny-related conflicts to identify those traits whose evolution cannot be explained with a given phylogenetic tree, explaining the conflicts as resulting from contact (Minett and Wang 2003, Nakhleh et al. 2005, Nelson-Sathi et al. 2011). The second class uses techniques for automatic sequence comparison to search for similar words across unrelated languages (van der Ark et al. 2007, Mennecier et al. 2016), related languages (Willems et al. 2016), or groups of related and unrelated languages (Hantgan and List forthcoming). The third class searches for distribution-related conflicts by comparing the amount of shared words within sublists of differing degrees of borrowability (Mc Mahon et al. 2005, Wang 2006, Galucio et al. 2015).

4.1 Phylogeny-Based Approaches to Borrowing Detection

The basic idea behind all phylogeny-based approaches to borrowing detection is that traits that are truly cognate should evolve without conflict along the true phylogeny of a given language family. If traits are in conflict with the phylogeny, this is assumed to be a direct hint for those traits to be borrowed. Consequently, this also means that the traits which were assumed to be cognate were wrongly annotated when creating the dataset. As an example, consider Figure 2, which displays two scenarios for the evolution of two characters along a phylogeny with four extant languages. In the first scenario, each character originates only once on the tree, and none of the characters is thus in conflict with the given phylogeny. In the second scenario, however, one character originates two times, on different branches. Since true cognate characters should only originate once on a phylogeny, this character is in conflict with the given phylogeny, and we can further assume that this conflict arises from contact among the languages in which this character is reflected.

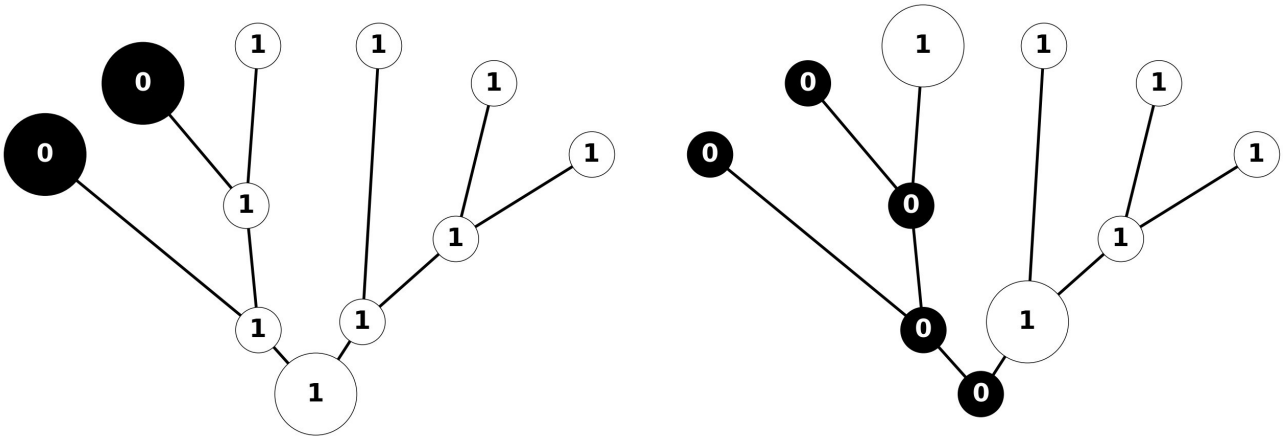

Figure 2: Two identical scenarios, the first involving no borrowing, the second involving a potential borrowing.

Different approaches to employ this originally biological technique of *character mapping* (Nunn 2011: 59) or *gain-loss mapping* (Cohen 2008) in the past, but the core of all approaches is to identify conflicts between a set of characters (cognates, structural traits) and their supposed evolution along a given reference phylogeny. The differences can be found in the data to which this method can be applied, the techniques being used to infer the different scenarios of character evolution, and in the way in which inferred conflicts are further analyzed and displayed.

Thus, in a first study known to me, Minett and Wang (2003) apply techniques of classical (unweighted) parsimony (Fitch 1971) to lexical cognates sets distributed over seven Chinese dialects to infer which of these cognate sets are in conflict with a given phylogenetic tree. In contrast to later approaches that binarize lexical cognates in phylogenetic analyses, their approach uses a multi-state modeling of lexical cognates (see List 2016 for a detailed discussion of different coding techniques) that does not allow that one concept is expressed by two synonymous words at the same time.

Another early study by Nakhleh et al. (2005) employs the same idea of searching for incompatible characters. In a second stage, the method tries to resolve the incompatible by turning the reference phylogeny into a phylogenetic network, in which vertical edges are added to the reference phylogeny. Similar ot the method by Minett and Wang, this approach also did not allow that one character is expressed by two different states in the same language. The method was tested on a dataset of lexical cognates, sound change processes, and morphological features across ancient Indo-European languages, coded as multi-state characters. According to the description provided by the authors, the linguistic data was heavily edited, requiring a very detailed historical knowledge about the languages under question. As a result, it seems that the method should only be applied as a tool for exploratory data analysis, where the knowledge about a language families history is in its initial stages.

Originally introduced as a method for the detection of gene transfer events (Dagan and Martin 2007), the *minimal lateral network* for automatic borrowing detection was applied in a couple of different studies and on different datasets, including Indo-European (Nelson-Sathi et al. 2011, List et al. 2014a), Chinese dialects (List et al. 2014b, List 2015), and Austronesian (Jäger and List 2018. The method also requires a reference phylogeny to be provided upon input and uses weighted parsimony applied to binary character states to infer which characters conflict with a given phylogeny. In contrast to alternative approaches, however, the minimal lateral network approach compares several scenarios with different weights for gain and loss events, using the *basic vocabulary size* criterion to select a weight ratio in which the number of words for ancestral languages is similar to the number of words in attested languages (List et al. 2014b.

That the method of mapping character evolution on trees and searching for those characters that are in conflict with the respective reference phylogenies is not only applicable to wordlists, but also to structural data, is shown in the recent work by Cathcart et al. (2018), in which the authors use a Bayesian likelihood framework for character mapping to identify areally transmitted traits from structural data for Indo-European languages.

Method	Reference	Characters	Algorithm	Data	Note
character-based borrowing detection	Minett and Wang 2003	multi-state	Fitch parsimony	lexical cognates, no synonyms	no code submitted
perfect phylogenetic networks	Nakhleh et al. 2005	multi-state	perfect phylogenies	lexical and structural cognates, no synonyms, data heavily edited	no code submitted
minimal lateral networks	Nelson-Sathi et al. 2011, List et al. 2014ab, List 2015	binary state	weighted parsiomony	basic vocabulary size as criterion for weight selection	
areal pressure	Cathcart et al. 2018	binary state	Bayesian likelihood framework	structural data	two-step procedure to infer areally transmitted traits

Table 1: Summary on phylogeny-based methods for borrowing detection.

Table 1 provides a summary comparison of the four different methods compared so far, including the data to which they were applied, the methods employed, and the literature in which the methods were applied so far. While methods based on character-mapping (or ancestral state reconstruction in general) have the advantage of being rather straightforward in their application, they suffer as well from a range of disadvantages. First, they require the phylogeny of the languages under question to be known in advance. Second, since not all characters that seem to conflict when mapped onto a reference phylogeny conflict indeed with it, given that parallel evolution is difficult to be ruled out, the methods tend to infer more borrowigns in the dateset than there are. Third, since the methods require a phylogeny, it is impossible to search for borrowings from or to unrelated languages.

As a final problem, given the lack of suitable test sets, it is difficult for us to estimate how well these methods work in the end. Given that the samples are often large, but that traditional methods do not provide us with detailed accounts on known borrowings for the languages under consideration, testing the success of a method against a gold standard of "known borrowings" is usually not feasible, and only reporting how many known borrowings the methods uncovered is likewise not enough. The only way to learn more about the method is an evaluation of the accuracy of the haracter-mapping techniques in general. Here, however, a recent study on ancestral state reconstruction techniques by Jäger and List (2018) shows not only that parsimony-based methods like MLN lag behind likelihood-based methods, but that the currently available test data itself suffers from many inconsistencies. This is also confirmed by the study of Cathcart et al. (2018), whose tests on the performance of their approach also show a rather low accuracy, especially when compared to other automatic methods like cognate detection or phonetic alignments (List 2014).

In order to enhance character mapping and phylogeny-based techniques for automatic borrowing

detection, it seems inevitable to invest more time in producing high-quality datasets for testing and training. It may be interesting in this context to note, however, that we can find similar problems in evolutionary biology, where techniques for the detection of lateral gene transfer are barely evaluated, and results vary greatly (Dagan and Martin 2006).

4.2 Sequence-Based Approaches to Borrowing Detection

While phylogeny-based approaches to the detection of lexical and structural borrowing draw their evidence for language contact exclusively from the topology of a given phylogenetic tree and the conflicts that arise when comparing the detailed evolution of lexical or structural traits, a second group of methods that have been proposed in the past take *word similarities* as their primary evidence. Given that words and morphemes can be easily modeled as *sequences of sounds* in historical linguistics (List 2014), it is possible to use techniques that were originally designed for sequence comparison in computer science and evolutionary biology to automatically determine the similarity of words across large datasets.

Although the techniques for automatic word comparison may differ considerably in their implementation and underlying "philosophy", the most successful methods that were proposed so far (Kondrak 2000, Nerbonne et al. 2011, Jäger 2013, List et al. 2018) all make use of techniques for *automatic sequence alignment* whose origins go back to the 1970s (Needleman and Wunsch 1970, Wagner and Fischer 1974, Smith and Waterman 1981). An alignment, in this context, is a specific technique by which two or more sequences are arranged in a matrix in such a way that corresponding segments appear in the same column, while segments for which no counterpart can be identified are contronted with gap symbols (List 2014). Once an alignment has been computed, it can be scored by comparing in how many columns of the matrix two or more sequences show different segments.

Alignment analyses are a very straightforward way to compute distance or similarity scores between sequence pairs. If words are presented in form of phonetic transcriptions, we can accordingly compute their phonetic distance, which can also be additionally informed by external knowledge on pronunciation similarities or sound change tendencies, yielding distance scores that try to mimick the way in which trained linguists would intuitively judge the similarity or difference of words drawn from different languages.

Given that we know well that sound change processes may yield words that look very different on the surface, but reflect the result of regular processes in their deeper structural similarity, naive approaches to measure phonetic differences with help of alignment analyses may easily fail to detect these "genotypic" (as opposed to "phenotypic", see Lass 1997) similarities. In order to overcome this problem, methods that search for regular sound correspondences across the languages under investigation, which are then incorporated into the distance calculation, have proven successful when searching for cognates across multi-lingual datasets (List 2012, List et al. 2017).

For the purpose of identifying borrowings, however, methods that measure only the surface similarity of words, have proven more useful, given that -- in contrast to regularly inherited words -- lexical borrowings show a high degree of surface similarity with their originals from which they were copied

into the recipient language. When comparing word similarities across genetically unrelated languages, as first proposed by van der Ark et al. (2007), surface similarities across words alone can serve as a proxy for the detection of borrowings. Note that the conflict in such an approach to automatic borrowing detection is still *phylogeny-related*, since the conflict observed in this approach is that of similarities observed for unrelated languages.

As pointed out by follow-up studies by Mennecier et al. (2016) and Zhang et al. (forthcoming), the importance of this sequence-similarity-based approach to the detection of lexical borrowings is the threshold, i.e., the cut-off point, by which a distance or similarity score computed by an alignment method is judged to be high (or low) enough to count as a potential borrowing. In order to determine these thresholds, annotated data is needed, i.e., data in which linguists have marked which words they consider as obvious borrowings. While the studies by Van der Ark (2007) and Mennecier et al. (2016) did not test the performance of different methods for phonetic alignment against each other, Zhang et al. (forthcoming) show that the rather simple, historically informed Sound-Class-Based Alignment (SCA) approach (List 2012b) largely outperforms earlier approaches, such as Heeringa's (2004) modified edit distance algorithm, or Wieling et al.'s (2012) rather sophisticated PMI-based scoring system derived from the data under investigation itself. With an accuracy of 85% achieved by List's SCA approach, this study also shows that the detection of recent borrowings with help fo surface sequence similarity measures is a most promising endeavor that should ideally be tested on more datasets in the future.

Two more recent studies expand on the rather simple idea of taking sequence similarities across unrelated languages as a proxy for the detection of lexical borrowings. Boc et al. (2010) and Willems et al. (2016) combine phylogenetic analysis with sequence similarity by computing individual *word trees* from pairwise word distances for all concepts across a given concept lists. These word trees are then analysed by *reconciling* each of them with a supposed (and user-determined) reference (or language) tree. The technique used for this comparison of a word tree with a reference tree is a rather popular technique from evolutionary biology where it is used to detect lateral gene transfer events. As of today, many different methods and models have been proposed, and it would go far beyond the scope of this chapter to discuss them in detail.

The algorithm used by Boc et al. and Willems et al., however, was first proposed by Makarenkov et al. (2006), and can be considered as outdated, especially given that no sufficient testing or discussion regarding the direct transfer of word distances to gene sequence distances in biology was carried out. The advantage, however, is that the methods are by now implemented as part of a larger web server package, called T-REX (http://www.trex.uqam.ca/, Boc et al. 2012), which makes it possible for users versed in data manipulation to easily test the methods themselves.

As a last method employing sequence similarities as its primary evidence, Hantgan and List (forthcoming) infer potential borrowings across related and unrelated languages by comparing word similarities derived from surface comparison ("phenotypic similarities") with word similarities based on automatically inferred language-specific sound correspondence probabilities ("genotypic similarities"), the former being represented by the aforementioned SCA algorithm for phonetic

alignment, an the latter being represented by the LexStat method for automatic cognate detection (List 2012). By searching explicitly for words that are similar according to their "phenotype" and ruling out words that are similar both "phenotypically" and "genotypically", they derive shared cognate percentages reflecting both a potentially borrowed and a potentially inherited layer. The application of this methods to a larger dataset including Dogon, Atlantic, Mande, Songhai, and Bangime, a language isolate, this methods confirms that the lexicon of Bangime is heavily influenced by the Dogon languages, but rather reflecting language contact than deeper genetic inheritance.

Similar to phylogeny-based approaches to borrowing detection, sequence-based approaches suffer from a series of shortcomings. The first problem, which can be observed for all approaches to borrowing detection, is the lack of suitable gold standard sets for testing and training of new methods. The second problem is the limited scope of most of these methods that allows their application to unrelated (Van der Ark 2007, Mennecier 2016) or related languages (Boc et al. 2010, Willems et al. 2016) only. As a third problem, sequence-based approaches are limited to lexical borrowings, and structural borrowings cannot be inferred with this method.

4.3 Borrowability-Accounts on Borrowing Detection

The last class of automatic approaches to handling language contact discussed in this chapter is also the oldest class of approaches. The idea that lexical concepts could be ranked by the expected borrowability of their counterparts in human languages was most prominently proposed by Swadesh (1950, 1952, 1955), but even in the work of Antoine Meillet (1866-1936) we can find statements emphasizing that certain concepts tend to be more stable and less prone to borrowing (Meillet 1921).

The idea, that concepts can be ranked by their relative borrowability, however, does not provide a concrete method to determine whether similar words are borrowings are not. While borrowability is regularly employed in classical approaches to studying language contact, an automatization requires a more formalized procedure.

The first to define such a procedure was (to my knowledge) Sergey Yakhontov (1926-2018), who proposed to divide a concept list of 100 concepts into a stable and a less stable part. Whenever the proportion of shared words between two or more languages would be higher (or similar) in the stable compared to the unstable sublist, thhe would take this as evidence for deeper genetic relationship. If the proportion showed the opposite behavior, with few words in the stable and many shared words in the unstable part, this was taken as evidence for contact. Although Yakhontov never published any study about this idea, his principle was employed by many colleagues, especially those affiliated with the Moscow School of Historical Linguistics. It is also in the work of his colleagues, especially that of Sergei Starostin (1953-2005), where we find the procedure described in due detail (Starostin 1991).

While it is difficult to say whether Yakhontov's idea of comparing two sublists with each other can be seen as an automatic procedure, it is clear that this idea is easy to formalize and automatize. Interestingly, the idea itself was later re-invented independently by a couple of scholars from different backgrounds. Thus, Chén (1996) proposed the same principle, but used different sublists of 100 items each as basis, to resolve questions of language contact in South East Asia. Chén's principle was then

also used to study the affiliation of the Bai language, concluding that Bai belongs to the Sinitic branch (Wang 2006), a question, that is still unresolved up to today (see for example Lee and Sagart 2008).

With tools for data-display, especially Neighbor-Net (Bryant and Moulton 2004), becoming more and more popular in historical linguistics and linguistic typology, scholars also started to test their suitability to study language contact. But since data-display networks cannot provide any hints regardig concrete processes, another principle was needed to differentiate between contact and inheritance. Here, McMahon and McMahon (2005), and McMahon et al. (2005), re-invented Yakhontov's sublist principle a third time, but while Yakhontov and Chén had divided one list into two, McMahon et al. derived two very small list from a big one, one stable list, labelled as "hihi" list, and one unstable list, labelled as "lolo". By computing Neighbor-Nets from the lexical distances derived from the two sublists, they seek to identify borrowings between different language groups by comparing the different networks. Unfortunately, the procedure is not further formalized, and while it offers a visualization that may have impressed scholars during that time, the real use of this procedure compared to the sublist approaches proposed by Yakhontov and Chén years before is questionable. The limited added value of using Neighbor-Nets this tasks may also reflect why the method was only followed up by a few scholars (for an exception, see Galucio et al. 2015).

The future needs to show whether approaches based on sublists of items prone and resistant to borrowing can provide new insights into language contact situations. Given attempts to propose a general scale of borrowability (Aikhenvald 2007), it might even be possible to employ borrowability arguments to study cases of language contact beyond the lexicon. For the time being, however, the methods have not been sufficiently tested, and further research is needed, especially to make sure that borrowability follows indeed general linguistic trends. For those interested in comparing the different concept lists discussed in this context, the Concepticon resource (List et al. 2016) provides a convenient way to compare them online, along with their original sources (https://concepticon.clld.org).

Outlook

It is quite possible that this overview will disappoint most readers interested in recent automatic approaches in historical linguistics and linguistic typology. While computational methods have enjoyed a great popularity of late in both disciplines, yielding promising results in phylogenetic reconstruction, cognate detection, and similar tasks, the development of methods for automatic borrowing detection, or automatic approaches to study language contact more broadly is still in their infancy. Some scholars may think that the slow progress in the automatic study of language contact represents the general tendency of scholars to prefer trees over networks and waves (Geisler and List 2013, Jacques and List forthcoming). It seems, however, that the problem is rooted more deeply.

By contrasting the classical methods for borrowing detection with the automatic methods that have been proposed so far, we can see that the problem lies in the practice of classical historical linguistics and classical linguistic typology itself. Neither of the two fields has proposed clear-cut procedures to distinguish genealogy from contact. While the comparative method for historical language comparison is usually praised to be a unique success story, no visible attempts to further formalize or advance the

method have been made. Handbooks on historical linguistics still treat the problem of detecting borrowings as one that is easily ruled out thanks to the comparative method, or one that is not worth being studied in isolation.

It is not clear whether the heuristics I mentioned in Section 3 are exhaustive. What is clear, howver, is that the current automatic methods have not yet exhausted their full potential. Of the three classes of methods presented in this chapter, the first two deal with phylogeny-related conflicts in the shared traits, while the last approach deals with borrowability and distribution-related conflicts. Trait-based conflicts, especially a more systematic treatment of recurrent sound correspondences and apparent conflicts within correspondence patterns, have not yet been studied automatically.

There is, thus, a lot to do, both for classical and for computational linguists. If we want to learn more about the cross-linguistic tendencies of language contact in all domains of language, we need to find ways to automatize at least some of the procedures we use, since the increasing amounts of data cannot be handled by classical methods alone. When designing automatic methods, however, it is important to keep a close eye on the classical approaches. Methods applied to linguistic data should never be blindly transferred from approaches employed in other scientific fields, but always be carefully adapted to our needs.

References

Aikhenvald, A. (2007): Grammars in contact. A cross-linguistic perspective. In: Aikhenvald, A. and R. Dixon (eds.): Grammars in contact. Oxford University Press: Oxford. 1-66.

van der Ark, R., P. Mennecier, J. Nerbonne, and F. Manni (2007): Preliminary identification of language groups and loan words in Central Asia. In: Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons. 13-20.

Baxter, W. and L. Sagart (2014): Old Chinese. A new reconstruction. Oxford University Press: Oxford.

Berg, T. (1998): Linguistic structure and change: An explanation from language processing. Clarendon Press: Gloucestershire.

Boc, A., A. Di Sciullo, and V. Makarenkov (2010): Classification of the Indo-European languages using a phylogenetic network approach. In: Locarek-Junge, H. and C. Weihs (eds.): Classification as a tool fo research. Springer: Berlin and Heidelberg. 647-655.

Boc, A., A. Diallo, and V. Makarenkov (2012): T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res.* 40.Web Server issue. W573-579.

Bowern, C. (2010): Historical linguistics in Australia: trees, networks and their implications. *Philosophical Transactions of the Royal Society B* 365.1559. 3845-3854.

Brown, C., E. Holman, and S. Wichmann (2013): Sound correspondences in the world's languages. *Language* 89.1. 4-29.

Bryant, D. and V. Moulton (2004): Neighbor-Net. An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21.2. 255-265.

Bryant, D., F. Filimon, and R. Gray (2005): Untangling our past: Languages, Trees, Splits and Networks. In: Mace, R., C. Holden, and S. Shennan (eds.): The evolution of cultural diversity: A phylogenetic approach. UCL Press: London. 67-84.

Cathcart, C., G. Carling, F. Larson, R. Johansson, and E. Round (2018): Areal pressure in grammatical evolution. An Indo-European case study. *Diachronica* 35.1. 1-34.

Chang, W., C. Cathcart, D. Hall, and A. Garret (2015): Ancestry-constrained phylogenetic analysis ssupport the Indo-European steppe hypothesis. *Language* 91.1. 194-244.

Chén Bǎoyà \hana 陈保亚 (1996): Lùn yǔyán jiēchù yǔ yǔyán liánméng \hana 论语言接触与语言联盟 [Language contact and language unions]. Yǔwén \hana 语文: Běijīng \hana 北京.

Běijīng Dàxué 北京大学 (1964): Hànyǔ fāngyán cíhuì [Chinese dialect vocabularies]. Wénzì Gǎigé 文字改革:

Cohen, O., N. Rubinstein, A. Stern, U. Gophna, and T. Pupko (2008): A likelihood framework to analyse phyletic patterns. *Philosophical Transactions of the Royal Society B* . .

Dagan, T. and W. Martin (2006): The tree of one percent. *Genome Biology* 7.118. 1-7.

Dagan, T. and W. Martin (2007): Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proceedings of the National Academy of Sciences* 104.3. 870-875.

Daval-Markussen, A. and P. Bakker (2011): A phylogenetic networks approach to the classification of English-based Atlantic creoles. *English World-Wide* 32.2. 115-136.

Dellert, J. (2016): Using causal inference to detect directional tendencies in semantic evolution. In: The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11).

Dellert, J. and G. Jäger (2017): NorthEuraLex (Version 0.9). Eberhard-Karls University Tübingen: Tübingen.

(2013): WALS Online. Max Planck Institute for Evolutionary Anthropology: Leipzig.

Eger, S. and A. Mehler (2016): On the linearity of semantic change. Investigating meaning variation via dynamic graph models. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics 52–58.

Fitch, W. (1971): Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology* 20.4. 406-416.

Friedman, V. (2007): Balkanizing the Balkan Sprachbund. In: Aikhenvald, A. and R. Dixon (eds.): Grammars in contact: A cross-linguistic typology.4. Oxford University Press: Oxford. 201-219.

Galucio, A., S. Meira, J. Birchall, D. Moore, N. Gabas Júnior, S. Drude, L. Storto, G. Picanço, and C. Rodrigues (2015): Genealogical relations and lexical distances within the Tupian linguistic family. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas* 10. 229-274.

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

Gray, R. and Q. Atkinson (2003): Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426.6965. 435-439.

Greenhill, S., R. Blust, and R. Gray (2008): The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4. 271-283.

Ben Hamed, M. and F. Wang (2006): Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23. 29-60.

Hantgan, A. and J.-M. List (forthcoming): Bangime: Secret language, language isolate, or language island?. *Journal of Language Contact* 0.0. .

Haspelmath, M. (2010): Comparative concepts and descriptive categories. *Language* 86.3. 663-687.

wrong type "phdthesis" {Heeringa2004}

Heggarty, P., W. Maguire, and A. McMahon (2010): Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B* 365.1559. 3829-3843.

Holman, E., C. Brown, S. Wichmann, A. Müller, V. Velupillai, H. Hammarström, S. Sauppe, H. Jung, D. Bakker, P. Brown, O. Belyaev, M. Urban, R. Mailhammer, J.-M. List, and D. Egorov (2011): Automated dating of the world's language families based on lexical similarity. *Current Anthropology* 52.6. 841-875.

Huson, D. (1998): SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14.1. 68-73.

Jacques, G. and J.-M. List (forthcoming): Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them). *Journal of Historical Linguistics* 19.1. ??-??.

Jäger, G. (2013): Phylogenetic inference from word lists using weighted alignment with empirical determined weights. *Language Dynamics and Change* 3.2. 245-291.

Jäger, G. (2018): Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data* 5.180189. 1-16.

Jakobson, R. (1960): Why 'Mama' and 'Papa'?. In: Perspectives in psychological theory: Essays in honor of Heinz Werner. 124-134.

Jarceva, V. (1990): . Sovetskaja Enciklopedija: Moscow.

Kaiping, G. and M. Klamer (2018): LexiRumah: An online lexical database of the Lesser Sunda Islands. *PLOS ONE* 13.10.

1-29.
Key, M. and B. Comrie (2016): The intercontinental dictionary series. Max Planck Institute for Evolutionary Anthropology: Leipzig.
Kluge, F. (2002): Etymologisches Wörterbuch der deut-schen Sprache . de Gruyter: Berlin.
Kondrak, G. (2000): A new algorithm for the alignment of phonetic sequences. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. 288-295.
Kondrak, G. (2002): Determining Recurrent Sound Correspondences by Inducing Translation Models. In: Nineteenth International Conference on Computational Linguistics (COLING 2002). 488-494.
Lass, R. (1997): Historical linguistics and language change. Cambridge University Press: Cambridge.
Lee, Y.-J. and L. Sagart (2008): No limits to borrowing: The case of Bai and Chinese. *Diachronica* 25.3. 357-385.
List, J.-M. (2012): Multiple sequence alignment in historical linguistics. A sound class based approach. In: Proceedings of ConSOLE XIX. 241-260.
List, J.-M. (2012): LexStat. Automatic detection of cognates in multilingual wordlists. In: Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources. 117-125.
List, J.-M., A. Terhalle, and M. Urban (2013): Using network approaches to enhance the analysis of cross-linguistic polysemies. In: Proceedings of the 10th International Conference on Computational Semantics -- Short Papers. Association for Computational Linguistics 347-353.
List, J.-M., S. Nelson-Sathi, H. Geisler, and W. Martin (2014): Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *Bioessays* 36.2. 141-150.
List, J.-M., S. Nelson-Sathi, W. Martin, and H. Geisler (2014): Using phylogenetic networks to model Chinese dialect history. *Language Dynamics and Change* 4.2. 222–252.
List, J.-M. (2014): Sequence comparison in historical linguistics. Düsseldorf University Press: Düsseldorf.
List, J.-M., T. Mayer, A. Terhalle, and M. Urban (eds.) (2014): CLICS: Database of Cross-Linguistic Colexifications. Version 1.0. Forschungszentrum Deutscher Sprachatlas: Marburg. http://clics.lingpy.org. http://www.webcitation.org/6ccEMrZYM.
List, J.-M. (2015): Network perspectives on Chinese dialect history. *Bulletin of Chinese Linguistics* 8. 42-67.
List, J.-M., M. Cysouw, and R. Forkel (2016): Concepticon. A resource for the linking of concept lists. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation. 2393-2400.
List, J.-M. (2016): Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1.2. 119-136.
List, J.-M., S. Greenhill, and R. Gray (2017): The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12.1. 1-18.
List, J.-M. (2018): Automatic inference of sound correspondence patterns across multiple languages. *bioRxiv* .434621. 1-25. [Preprint, under review, not peer-reviewed]
List, J.-M., M. Walworth, S. Greenhill, T. Tresoldi, and R. Forkel (2018): Sequence comparison in computational historical linguistics. *Journal of Language Evolution* 3.2. 130–144.
List, J.-M., S. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel (2018): CLICS². An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats. *Linguistic Typology* 22.2. 277-306.
刘俐李, L., W. 王洪钟, and B. 柏莹 (2007): Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí 现代汉语方言核心词·特征词集 [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]. Fènghuáng 凤凰: Nánjīng 南京.
Maddieson, I., S. Flavier, E. Marsico, C. Coupé, and F. Pellegrino. (2013): LAPSyD: Lyon-Albuquerque Phonological Systems Database. In: Proceedings of Interspeech.
Makarenkov, V., A. Boc, C. Delwiche, A. Diallo, and H. Philippe (2006): New Efficient Algorithm for Modeling Partial and Complete Gene Transfer Scenarios. In: Data Science and Classification. 341-349.
Malkiel, Y. (1954): Etymology and the Structure of Word Families. *Word* 10.2-3. 265-274.
Matisoff, J. (2015): The Sino-Tibetan Etymological Dictionary and Thesaurus project. University of California: Berkeley.

McMahon, A. and R. McMahon (2005): Language classification by numbers. Oxford University Press: Oxford.

McMahon, A., P. Heggarty, R. McMahon, and N. Slaska (2005): Swadesh sublists and the benefits of borrowing: An Andean case study. *Transactions of the Philological Society* 103. 147-170.

McMahon, A., P. Heggarty, R. McMahon, and W. Maguire (2007): The sound patterns of Englishes. Representing phonetic similarity. *English Language and Linguistics* 11.1. 113-142.

Meier-Brügger, M. (2002): Indogermanische Sprachwissenschaft. de Gruyter: Berlin and New York.

Meillet, Antoine. 1921[1965]. Linguistique historique et linguistique générale. Paris: Libr. Champion.

Mennecier, P., J. Nerbonne, E. Heyer, and F. Manni (2016): A Central Asian language survey. *Language Dynamics and Change* 6.1. 57–98.

Minett, J. and W.-Y. Wang (2003): On detecting borrowing. *Diachronica* 20.2. 289–330.

Moran, S., D. McCloy, and R. Wright (eds.) (2014): PHOIBLE Online. Max Planck Institute for Evolutionary Anthropology: Leipzig. http://phoible.org/.

Morrison, D. (2014): Phylogenetic networks: a new form of multivariate data summary for data mining and exploratory data analysis. *WIREs Data Mining and Knowledge Discovery* . .

Nakhleh, L., D. Ringe, and T. Warnow (2005): Perfect Phylogenetic Networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81.2. 382-420.

Needleman, S. and C. Wunsch (1970): A gene method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48. 443-453.

Nelson-Sathi, S., J.-M. List, H. Geisler, H. Fangerau, R. Gray, W. Martin, and T. Dagan (2011): Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society of London B: Biological Sciences* 278.1713. 1794-1803.

Nerbonne, J., R. Colen, C. Gooskens, P. Kleiweg, and T. Leinonen (2011): Gabmap -- A web application for dialectology. *Dialectologia* Special Issue II. 65-89.

Nunn, C. (2011): The comparative approach in evolutionary anthropology and biology. University of Chicago Press: Chicago and London.

Pfeifer, W. (1993): Etymologisches Wörterbuch des Deutschen . Akademie: Berlin.

Polyakov, V. and V. Solovyev (2006): Kompjuternye modeli i metody v tipologii i komparativistike [Computational models and methods in typology and comparative linguistics]. Kazan'skij universitet: Kazan'.

Prokić, J., M. Wieling, and J. Nerbonne (2009): Multiple sequence alignments in linguistics. In: Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education. 18-25.

Prokić, J. (2010): Families and resemblences . . Rijksuniversiteit Groningen: Groningen.

Smith, T. and M. Waterman (1981): Identification of common molecular subsequences. *Journal of Molecular Biology* 1. 195-197.

Starostin, S. (1991): Altajskaja problema i proischo\vzdenije japonskogo jazyka [The Altaic problem and the origin of the Japanese language]. Nauka: Moscow.

Starostin, S. (1995): Old Chinese vocabulary: A historical perspective. In: Wang, W.-Y. (ed.): The ancestry of the Chinese language. University of California Press: Berkeley. 225-251.

Starostin, G. (ed.) (2008): Tower of Babel. An etymological database project. http://starling.rinet.ru.

Steiner, L., P. Stadler, and M. Cysouw (2011): A pipeline for computational historical linguistics. *Language Dynamics and Change* 1.1. 89-127.

Sturtevant, E. (1920): The pronunciation of Greek and Latin. University of Chicago Press: Chicago.

Swadesh, M. (1950): Salish internal relationships. *International Journal of American Linguistics* 16.4. 157-167.

Swadesh, M. (1952): Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96.4. 452-463.

Swadesh, M. (1955): Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21.2. 121-137.

Szeto, P., U. Ansaldo, and S. Matthews (2018): Typological variation across Mandarin dialects: An areal perspective with a quantitative approach. *Linguistic Typology* 22.2. 233-275.

Tadmor, U. (2009): Loanwords in the world's languages. Findings and results. In: Haspelmath, M. and U. Tadmor (eds.): Loanwords in the world's languages. de Gruyter: Berlin and New York. 55-75.

Wagner, R. and M. Fischer (1974): The string-to-string correction problem. *Journal of the Association for Computing Machinery* 21.1. 168-173.

Wang, W.-Y. (2006): Yǔyán, yǔyīn yǔ jìshù \hana 語言,語音與技術 [Language, phonology and technology]. Xiānggǎng Chéngshì Dàxué: Shànghǎi 上海.

Whewell, W. (1847): The philosophy of the inductive sciences, fofound upon their history. John W. Parker: London.

Wieling, M., E. Margaretha, and J. Nerbonne (2012): Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics* 40.2. 307-314.

Willems, M., E. Lord, L. Laforest, G. Labelle, F.-J. Lapointe, A. Di Sciullo, and V. Makarenkov (2016): Using hybridization networks to retrace the evolution of Indo-European languages. *BMC Evolutionary Biology* 16.1. 1-18.

Willis, D. (2011): Reconstructing last week's weather: Syntactic reconstruction and Brythonic free relatives. *Journal of Linguistics* 47.2. 407-446.

Wilson, E. (1998): Consilience. The unity of knowledge. Vintage Books: New York.

Haspelmath, M. and U. Tadmor (eds.) (2009): World Loanword Database. Max Planck Digital Library: Munich. http://wold.livingsources.org.

Zenner, E., D. Speelman, and D. Geeraerts (2014): Core vocabulary, borrowability and entrenchment. *Diachronica* 31.1. 74–105.

Zhang, L., F. Manni, R. Fabri, and J. Nerbonne (under review): Detecting loan words computationally. In: : Draft, submitted to the Contact Language Libraries series. Benjamins: Amsterdam.