Authors final copy before production, contains the supplementary information as well. Rough page marks of published texts are given in square brackets for the main text. Please cite as:

List, Johann-Mattis (2019): Automated methods for the investigation of language contact, with a focus on lexical borrowing. Language and Linguistics Compass 13.e12355. 1-16. DOI: 10.1111/lnc3.12355.

Automated methods for the investigation of language contact, with a focus on lexical borrowing

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 automated 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 automated approaches. This study provides a concise introduction to the most important approaches which have been proposed in the past, presenting methods 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. While the overview focuses on approaches dealing with lexical borrowing, questions of general contact inference will also be discussed where applicable.

Keywords

computational historical linguistics, computational areal linguistics, automated contact inference, automated borrowing detection, language contact, sequence comparison, phylogenetic networks

1 Introduction

The past two decades have seen a drastic increase of quantitative applications in historical linguistics and linguistic typology, witnessed by multiple articles dealing with the automation of formerly exclusively manual tasks, such as phylogenetic reconstruction (Gray & Atkinson, 2003; Holman et al., 2011), word comparison (Kondrak, 2000; List, Walworth, et al., 2018; Prokić et al., 2009), semantic change (Dellert, 2016; Eger & Mehle, 2016; Steiner et al., 2011), and regular sound correspondences (Brown et al., 2013; Kondrak, 2002; List, 2019). The *quantitative turn* was specifically favored by the compilation of large databases, offering cross-linguistic accounts on typological structures (Dryer & Haspelmath, 2013; Polyakov & Solovyev, 2006), lexical cognates (Greenhill et al., 2008; Matisoff, 2015; Starostin, 2008), lexical data in general (Dellert & Jäger, 2017; Kaiping & Klamer, 2018), phoneme inventories (Maddieson et al., 2013; Moran et al., 2014), and polysemies (List, Greenhill, et al., 2018). [1]

Given the importance of language contact for the study of language history and linguistic typology, it is not surprising that automated approaches to study language contact were also proposed. In contrast to 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 (Bryant & Moulton, 2004), as implemented by the SplitsTree software package (Huson, 1998). SplitsTree can be conveniently used with various data types, including lexical

data (Ben Hamed & Wang, 2006; Bryant et al., 2005), phonetic data (Heggarty et al., 2010; McMahon et al., 2007; Prokić, 2010), and typological data (Daval-Markussen & Bakker, 2011; Szeto et al., 2018).

While splits networks are a useful way summarize a dataset for *exploratory data analysis* (Morrison, 2014), they do not allow to study language contact directly, because they do not allow to infer *which traits* have been shared as a result of contact. Since splits networks and similar approaches (like the numerous attempts to provide visualizations or numeric accounts of Schmidt's *Wave theory* from 1872) lack the time dimension, they do not allow us to infer historical processes by distinguishing borrowed from inherited traits (for details, compare Jacques & List 2019, pp. 138-142). Therefore, this overview will not discuss splits networks and similar approaches, but will instead focus on methods which allow for a concrete interpretation of findings by presenting explicit historical scenarios, or explicit instances of borrowing among different languages.

This overview will first look at the general problem of identifying contact-induced similarities between languages. We will then discuss how these problems are dealt with in non-automated frameworks. These classical approaches will then be contrasted with the most promising automated techniques that have been proposed so far, including phylogenetic reconstruction (Section 4.1), sequence comparison (Section 4.2), and borrowability scales (Section 4.3).

2 Similarities and language contact

No matter whether one is interested in inherited or borrowed traits, without resorting to some notion of *similarity* across languages, it is not possible to study historical language relations. Depending on what traits (*comparative concepts*, in the sense of Haspelmath (2010)) we inspect, languages can resemble in various ways. They can share similar words, but also similar structures. While some similarities may give us concrete hints regarding shared histories, many of the similarities we can observe are coincidental or based on general ("universal") tendencies 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, consider Modern Greek θεός [θɛɔs] 'god' and Spanish *dios* [diɔs] 'god'. Although both words look and sound similar, this is a coincidence, as we see from their oldest ancestors, Old Latin *deivos* and Mycenaean Greek *thehós* (Meier-Brügger, 2002, p. 57f). As an example for the second type, consider Chinese *māmā* 媽媽 'mother' vs. German *Mama* 'mother'. Both words are similar, but only because they reflect general principles of early language acquisition (Jakobson, 1960). An example for genealogical similarity are German *Zahn* and English *tooth*, both going back to Proto-Germanic **tanθ*-. Contact-induced similarity is reflected in English *mountain* and French *montagne*, with the former borrowed from the latter. [2]

Following List (2014), we can display these similarities in the decision tree shown in Figure 1. Here, the last two types of similarity are highlighted, indicating that they are *historical*, reflecting individual scenarios of language change. When searching for contact-induced similarities, it is crucial to distinguish between the four similarity types. At times this can be trivial, especially if we can rule out that languages sharing similar traits are related. In such cases, all that needs to be shown is that that similarities are unlikely to have evolved independently. Identifying the dynamics of language contact among genetically closely related languages, on the other hand, is much more difficult.

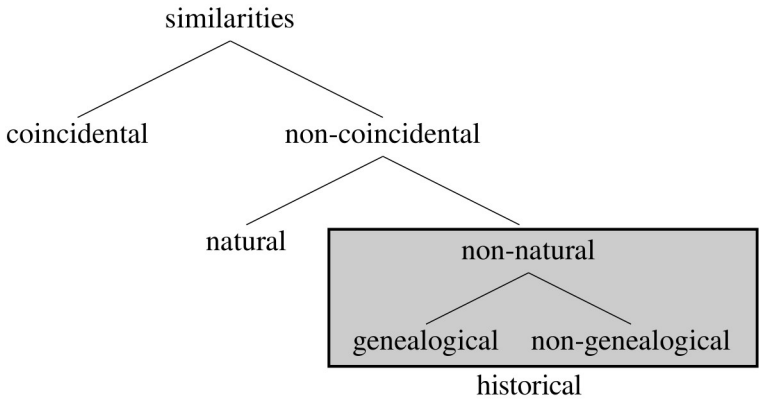


Figure 1: Reasons for similarities between languages.

3 Classical approaches to the study of contact situations

The most straightforward way to study language contact is by means of *direct evidence*. 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. While sources from the 1960s (Běijīng Dàxué, 1964) list only the form [jit²²tʰɐu²¹₃₅] 熱頭, more recent vocabulary collections list exclusively the former form (Liú et al., 2007). If languages are well-documented across time, we can often directly see when a word enters their lexicon. If there is no direct evidence, scholars need to resort to indirect techniques to prove that traits arose from contact. In contrast to general language change, contact-induced change does not proceed in a largely regular manner, but can be seen as a disruptive and chaotic event that *may* occur but might as well not occur during language history.

While historical linguistics has developed sophisticated techniques to prove that language similarities are genealogical, the techniques for identifying contact-induced similarities are less

homogenous, involving detailed sifting of multiple pieces which are only in combination convincing. In this regard, techniques for contact detection are not much different from other, more specific, types of linguistic reconstruction, such as the "philological reconstruction" of ancient pronunciations (Jarceva, 1990; Sturtevant, 1920), the reconstruction of detailed etymologies (Malkiel, 1954), or the reconstruction of syntax (Willis, 2011). Despite the difficulty in determining exact workflows, we can identify a couple of *proxies* that scholars use to assess whether a given trait has been 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. Since English *mountain* is reflected only in English, with similar words only in Romance, we could take this as evidence that the English word was borrowed. Since these conflicts 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 observed in the case of irregular sound correspondence patterns. German *Damm*, for example, is related to English *dam*, [3] 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: s. v. *Damm*), 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 can be derived from distributional properties of 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, religion, this is usually also seen as a strong indicator of borrowing, since we expect that related languages share words across different fields, including *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 as well. The lack of an infinitive in Balkan languages, such as Bulgarian and Greek (Friedman, 2007, p. 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, two further types of evidence are of great importance for contact inference. 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

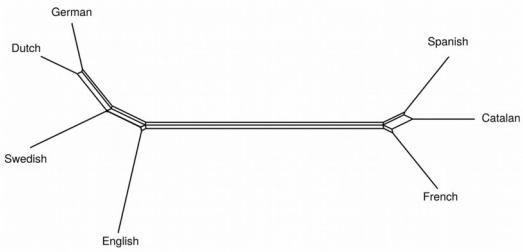
different languages, it is self-evident that areal proximity, including proximity by means of travel routes, is a necessary argument when proposing contact relations between different varieties.
Since direct evidence confirms that linguistic interference does not act to the same degree on all levels of linguistic organisation, the notion of *borrowability* also plays an important role. Although scholars tend to have different opinions about the concept, most would probably agree with the borrowability scale proposed by Aikhenvald (2007b, p. 5), which ranges from "inflectional morphology" and "core vocabulary", representing aspects resistant to borrowing, up to "discourse structure" and the "structure of idioms", representing aspects easy to borrow. How core vocabulary can be defined, and how the borrowability of individual concepts can be determined and ranked, however, has been subject to controversial debates (Lee & Sagart, 2008; Starostin, 1995; Tadmor, 2009; Zenner et al., 2014).

4 Computational approaches to study of language contact

Despite the large number of quantitative applications during the last two decades, computational approaches to infer contact situations are still in their infancy. As of now, none of the few approaches proposed so far can compete with the classical methods. The reasons for this are twofold. First, given the multiple types of evidence employed by the classical approaches, the formalization of the problem of borrowing detection is difficult. Second, given the limited number and suitability of datasets annotated for different types of linguistic interference, scholars have a hard time in developing algorithms, since they lack data for testing and training.

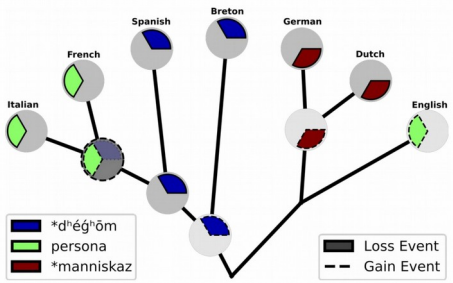
In principle, all algorithms for contact inference that have been proposed so far make use of the strategies used in the classical approaches. Thus, they infer or determine shared traits among two or more languages, and then determine conflicts in these traits, taking areal closeness and borrowability into account. In contrast to classical approaches, which combine different types of evidence, computational approaches are usually restricted to one type. [4]

The automated methods proposed so far can be divided into three 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 (Section 4.1). The second class uses techniques for automated sequence comparison to search for similar words but not cognate words across different languages (Section 4.2). The third class searches for distribution-related conflicts by comparing the amount of related words within sublists of differing degrees of borrowability (Section 4.3). Figure 2 contrasts these three different approaches with distance-based approaches such as splits networks, which are not further discussed here in some further detail, along with graphics for illustration. A short tutorial presenting how the different methods can be employed to search for borrowings in the same datasets is available in the supplementary material accompanying this paper.



When trying to infer a phylogenetic tree from distance data, methods such as the Neighbor-Net algorithm allow for a sophisticated display of shared characteristics among the taxonomic units under investigation. However, since they aggregate signal, they are not helpful at identifying concrete borrowing events.

A Distance-based methods



Based on a reference phylogeny and a dataset consisting of presumed cognate words, methods searching for phylogeny-related conflicts method assume that all cognates words which evolve independently on a phylogeny (as proposed by an automaed ancestral state reconstruction method) result from borrowings, such as in the case of words for 'person', where English, French, and Italian share the same word.

B Phylogeny-related conflicts

ID	DOCULECT	CONCEPT	SCAID	ALIGNMENT
2448	Breton	person	632	d ə: n
2450	Dutch	person	633	m ɛ n s
2444	German	person	633	m ɛ n ʃ
2460	Spanish	person	638	o m b r e
2445	Italian	person	634	p e r s o n a
2451	French	person	634	p ɛ ʁ s ɔ n -
2447	English	person	634	p ɜ: - s ə n -

Alignments are a basic way to compare words across different languages. If words in distantly related languages are more similar than expected, this may be a hint that one of the words showing high similarity must have been borrowed. The example shows an alignment of words for 'person' in Germanic and Romance languages, with English *person* being very similar to the words in the Romance languages.

C Similar words across unrelated languages

Concept	Sublist	Borrowed Score
ashes	unstable	0.15
dog	stable	0.16
drink	unstable	0.10
mountain	unstable	0.17
stone	stable	0.10
road	unstable	0.31

Scholars employ borrowability arguments specifically in those cases where additional evidence is lacking. Especially in those cases where it is not clear if traits are inherited or not, the use of sublists, which either rank items completely, or divide them into stable and unstable sets are used to support or discard arguments for or against the common descent of traits.

D Distribution-related conflicts

[6]

Figure 2: Comparing different approaches for automated contact inference, including distance-based methods.

4.1 Phylogeny-Based Approaches to Borrowing Detection

The basic idea behind all phylogeny-based approaches to borrowing detection is that truly cognate traits 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 that these traits were 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 the scenario for the evolution of words meaning 'human being' in Romance, Germanic, and Celtic languages in Figure 2B. While we find reflexes of Latin *persona* 'mask' in Italian and French (and also Spanish, but not in this particular dataset used for this example), we also find the word *person* in English. By inferring how the words most probably evolved along the given phylogeny, we can see a conflict involving the reflexes of Latin *persona*, as they evolve two times on the tree, one time in Romance, and one time in English. This conflict of the evolution of one character in the phylogeny can be interpreted as resulting from a borrowing event, and we know, of course, that this is true for the case of English *person*.

Different approaches to employ this originally biological technique of *character mapping* (Nunn, 2011, p. 59) or *gain-loss mapping* (Cohen et al., 2008) have been proposed 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 the first study of this type known to me, Minett & Wang (2003) apply techniques of classical (unweighted) parsimony (Fitch, 1971) to lexical cognate sets distributed over seven Chinese dialects to infer which of these cognate sets are in conflict with a given phylogeny. In contrast to later approaches which binarize lexical cognates, 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 one concept to be expressed by two synonymous words.

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 them by turning the reference phylogeny into a network, in which horizontal edges reflect contact. Similar to the method by Minett & Wang (2003), this approach also cannot handle synonymous entries. 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 preparation of the linguistic data required a very [5] detailed historical knowledge about the languages in question. Hence, the method should only be applied to languages whose history is well-known.

Originally introduced as a method for the detection of gene transfer events (Dagan & Martin, 2007), the *minimal lateral network* (MLN) method for automated borrowing detection was applied in a couple of different studies and on different datasets, including Indo-European (List, Nelson-Sathi, Geisler, et al., 2014; Nelson-Sathi et al., 2011), Chinese dialects (List, 2015; List, Nelson-Sathi, Martin, et al., 2014), and Austronesian (Jäger, 2018). MLN uses weighted parsimony applied to binary character states to infer which characters conflict with a given phylogeny. In contrast to alternative approaches, however, MLN 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, Nelson-Sathi, Martin, et al., 2014).

That character mapping techniques are not only applicable to wordlists, but also to structural data, is shown in the recent work by Cathcard et al. (2018), in which the authors use a Bayesian likelihood framework for character mapping to identify areally transmitted traits from structural data among Indo-European languages.

Method	character-based borrowing detection	perfect phylogenetic networks	minimal lateral networks	areal pressure in language evolution
Reference	Minett and Wang 2003	Nakhleh et al. 2005	Nelson-Sathi et al. 2011	Cathcart et al. 2018
Character model	multi-state	multi-state	binary	binary
Algorithm	Fitch parsimony	perfect phylogenies	weighted parsimony	Bayesian likelihood
Data	lexical cognates	lexical & structural cogn.	lexical cognates	structural data
Synonyms	not allowed	not allowed	allowed	not allowed
Code availability	-	-	+	+

Table 1: Comparing different character mapping techniques for the purpose of contact inference.

Table 1 provides a summary of the four different methods compared so far, including the data to which they were applied, the methods employed, the literature in which they were applied, and information on code availability. While character-mapping methods are rather straightforward in their application, they suffer 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 excluded, the methods tend to infer more borrowings in the dataset 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 to estimate how well character mapping works in the end. The samples are often too large to allow for a detailed comparison with traditional methods. Testing the success of a method against a gold standard of "known borrowings" is therefore not feasible. The only way to learn more about the methods is a general evaluation of the character-mapping techniques. Here, however, the study by Jäger (2018) shows not only that parsimony-based methods like MLN lag behind likelihood-based methods, but also that the currently available test data themselves suffer from inconsistencies. [7]

In order to enhance phylogeny-based techniques for automated contact inference, it seems inevitable that we must invest more time in producing high-quality datasets for testing and training. It is 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 & Martin, 2006).

4.2 Sequence-Based Approaches to Borrowing Detection

While phylogeny-based approaches to contact inference draw their evidence exclusively from the conflicts between reference phylogenies and individual trait evolution, another family of methods takes *word similarities* as primary evidence. Given that words can be easily modeled as *sequences of sounds* (List, 2014), it is possible to use techniques that were originally designed for sequence comparison in computer science and evolutionary biology to automatically determine word similarities across large datasets.

Although automated word comparison techniques differ in implementation and "philosophy", the most successful methods proposed so far (Jäger, 2013; Kondrak, 2000; List, Walworth, et al., 2018; Nerbonne et al., 2011) all make use of techniques for *automated sequence alignment* whose origins go back to the 1970s (Needleman & Wunsch, 1970; Smith & Waterman, 1981). An alignment is a specific technique by which sequences are arranged in a matrix in such a way that corresponding segments appear in the same column, while segments without a counterpart are confronted with gap symbols (List, 2014). Once alignments have been computed, word pairs can be scored for similarity by comparing in how many columns of the matrix the sequences differ.

Alignment analyses are a very straightforward way to compute distance scores between word pairs. If words are presented in form of phonetic transcriptions, they provide a *phonetic distance*, which can also be additionally informed by external knowledge on pronunciation similarities or sound change tendencies, yielding distance scores that try to mimic the way in which trained linguists would intuitively judge word similarity.

Given that we know well that sound change may yield words that look very different on the surface, but reflect 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" (Lass, 1997) – similarities. In order to overcome this problem, methods that first search for sound correspondences across languages and then incorporate them into the distance calculation have proven successful when searching for cognates across multilingual datasets (List et al., 2017, Dellert 2018).

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 the words from which they were copied into the recipient language. When comparing word similarities across unrelated languages, as first proposed by Ark et al. (2007), surface similarities alone can serve as a proxy for borrowing detection.

As shown in follow-up studies (Mennecier et al., 2016; Zhang et al., 2019), the cut-off point, or threshold, by which words are automatically judged to be similar or not is crucial for the success of sequence-based approaches to contact inference. In order to determine these thresholds, annotated data is needed, in which linguists have marked which words they consider as obvious borrowings. While the studies by Ark et al. (2007) and Mennecier et al. (2016) did not test the performance of different methods for phonetic alignment against each other, Zhang et al. (2019) show that the rather simple, historically informed Sound-Class-Based Alignment (SCA) approach (List, 2012) largely outperforms earlier approaches, such as the modified edit distance algorithm by Heeringa (2004), or the rather sophisticated PMI-based scoring system derived from the data under investigation itself by Wieling et al. (2012). [8]

Two more recent studies expand on the rather simple but straightforward idea of the approaches mentioned above. 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 list. These word trees are then analyzed by *reconciling* each of them with a reference phylogeny. This "tree reconciliation" technique is rather popular in 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 survey to discuss them in detail. Nakhleh (2013) is a very good starting point for interested readers to learn more about tree reconciliation techniques.

The algorithm used by Boc et al. and Willems et al., was first proposed by Makarenkov et al. (2006) for the purpose of tree reconciliation in biology. Unfortunately, the authors do not discuss to which degree this direct transfer from a biological algorithm applied to gene distances to word distances in linguistics is fruitful after all. In addition, neither of the studies provides rigorous

tests of the inferences, which reflects the general problem of lacking data for testing and training, already observed for phylogeny-based approaches to contact inference. An advantage of the approach by Makarenko et al. is that the methods have now been implemented as part of a larger web server package (Boc et al., 2012), which makes it possible for users to easily test the methods themselves.

As a last method employing sequence similarities as its primary evidence, Hantgan & List (forthcoming) infer potential borrowings across related and unrelated languages by comparing word similarities derived from surface comparisons ("phenotypic similarities"), with word similarities based on language-specific sound correspondence probabilities ("genotypic similarities", see Lass 1997, p. 130 for a distinction between the different kinds of similarities), the former being represented by the SCA algorithm, and the latter being represented by the LexStat method for automated 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. Applying the approach to a larger dataset including Dogon, Atlantic, Mande, Songhai, and the language isolate Bangime, they showed that the lexicon of Bangime is heavily influenced by Dogon languages, but reflects contact rather than inheritance.

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

4.3 Borrowability-Accounts on Borrowing Detection

The last class of automated 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 (Swadesh, 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 [9]
borrowing (Meillet, n.d.). The idea, that concepts can be ranked by their relative borrowability, however, does not provide a concrete method to determine borrowings. While borrowability is regularly employed in classical approaches to studying language contact, an automatization requires a formalized procedure.

The first to define such a procedure was (to my knowledge) Sergey Yakhontov (1926-2018),

who proposed to divide a concept list into a stable and a less stable part. Whenever the proportion of related words between two or more languages would be higher in the stable compared to the unstable sublist, he would take this as evidence for deeper genetic relationship. If the proportion showed the opposite behavior, with few words in the stable and many related 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, in whose work, especially that of Sergei Starostin (1953-2005), we find the procedure described in due detail (Starostin, 1991).

While it is difficult to say whether Yakhontov's idea of comparing sublists can be seen as an automated procedure, it is clear that this idea is easy to formalize and automate. Interestingly, the idea itself was later re-invented independently by scholars from different backgrounds. Thus, Chén (1996) proposed the same principle, but used different sublists to resolve questions of language contact in South East Asia. Chén's principle was then also used to study the affiliation of Bai (Wang, 2006), a question that is still unresolved up to today (Lee & Sagart, 2008).

With tools like Neighbor-Net (Bryant & Moulton, 2004) becoming more and more popular in diversity linguistics, scholars also started to test their suitability to study language contact. But since data-display networks cannot provide any hints regarding concrete processes, another principle was needed to differentiate between contact and inheritance. Here, A. McMahon & McMahon (2005) and A. 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 lists from a big one, a stable list, labelled as "hihi", and an unstable list, labelled as "lolo". By computing Neighbor-Nets from the lexical distances derived from the sublists, they tried to identify borrowings comparing the networks. Unfortunately, the procedure is not further formalized, and while it offers a visualization of differences between sublists, the real use of this procedure compared to the sublist approaches by Yakhontov and Chén is questionable, and the method was only sporadically followed up (Galucio et al., 2015).

The future will 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, 2007a), it might even be possible to employ borrowability arguments to study language contact beyond the lexicon (Nichols 2003). For the time being, however, the methods have not been sufficiently tested, and further research is needed, especially to make sure that borrowability follows general linguistic trends.[1]

1 For those interested in comparing the different concept lists discussed in this context, the Concepticon resource (List, Cysouw, et al., 2016) provides a convenient way to compare them online, along with their original sources.

5 Conclusion and Outlook

This overview may seem disappointing. While computational methods now enjoy a growing popularity in diversity linguistics, yielding promising results in phylogenetic reconstruction, cognate detection, and similar tasks, the development of methods for automated contact inference is still in its infancy. One might think that this represents the general tendency of scholars to prefer trees over networks and waves (Geisler & List, 2013, Jacques & List 2019). It seems, however, that the problem is rooted more deeply. [10]

By contrasting the classical methods for borrowing detection with the automated methods that have been proposed so far, we can see that the detection of borrowing is not only difficult for automated methods, but also for the classical disciplines of historical linguistics and linguistic typology itself. Classical ("manual") borrowing detection is based on the meticulous sifting of multiple pieces of evidence. Rather than mechanically applying a unified method, scholars make use of cumulative evidence to search for a scenario that explains the different kinds of data best. By nature, "cumulative-evidence arguments" (Berg, 1998) — arguments based on *consilience* (Whewell, 1847; Wilson, 1998) — are more difficult to formalize than clear-cut procedures that yield simple, binary results. Therefore, it is often very difficult to formalize what scholars do concretely in order to come to their conclusions. That scholars resort to cumulative evidence arguments follows to a large part from the phenomenon of language contact itself. While languages have been shown to change in a surprisingly regular manner in the absence of contact, the phenomenon of language contact is largely chaotic and may often show idiosyncratic patterns and pathways.

Although a further systematization of methods for contact inference – be they manual or automated– meets several obstacles, a broader discussion regarding the application range and usefulness of different heuristics that scholars have proposed so far would be generally desirable. Not only would it help young scholars in learning existing techniques, but it might also encourage scholars to discuss and develop new methods and techniques.

While the field would beyond doubt profit from a more systematic treatment of various techniques for contact inference, it is also obvious that the current automated methods have not yet exhausted their full potential. Of the three classes of automated methods presented here, the first two deal with phylogeny-related conflicts in shared traits, while the last approach deals with distribution-related conflicts and borrowability. Trait-based conflicts, especially a more systematic treatment of recurrent sound correspondences and apparent conflicts within correspondence patterns, have not yet been studied automatically.

What will be crucial for the further advancement of automated methods for contact inference in general is the availability of datasets for testing and training. The problem of creating such datasets, however, lies not only in the labor involved in digitizing and annotating data itself, but also in the limited knowledge we have with respect to borrowing processes in the world's languages. Since borrowing is generally hard to detect, not only for automated methods, it is even harder to do so exhaustively when trying to create a benchmark database for training and testing.

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 automate at least some of the procedures we use, since the increasing amounts of data cannot be handled by classical methods alone. When designing automated 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. [11]

References

Aikhenvald, A. Y. (2007a). Grammars in contact. A cross-linguistic perspective. In A. Y. Aikhenvald & R. M. W. Dixon (Eds.), *Grammars in contact* (pp. 1–66). Oxford: Oxford University Press.

Aikhenvald, A. Y. (2007b). Semantics and pragmatics of grammatical relations in the Vaups linguistic area. In A. Y. Aikhenvald & R. M. W. Dixon (Eds.), *Grammars in contact: A cross-linguistic typology* (Vol. 4, pp. 237–266). Oxford: Oxford University Press.

Ark, R. van der, Mennecier, P., Nerbonne, J., & Manni, F. (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* (pp. 13–20).

Běijīng Dàxué, 北. (1964). *Hànyǔ fāngyán cíhuì* 汉语方言词汇 [Chinese dialect vocabularies]. Běijīng: Wénzì Gǎigé.

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

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

Boc, A., Diallo, A. B., & Makarenkov, V. (2012). T-rex: A web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Research*, *40*, 573–579.

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

Brown, C. H., Holman, E. W., & Wichmann, S. (2013). Sound correspondences in the world's languages. *Language*, *89*(1), 4–29.

Bryant, D., Filimon, F., & Gray, R. D. (2005). Untangling our past: Languages, Trees, Splits and Networks. In R.

Mace, C. J. Holden, & S. Shennan (Eds.), *The evolution of cultural diversity: A phylogenetic approach* (pp. 67–84). London: UCL Press.

Bryant, D., & Moulton, V. (2004). Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, *21*(2), 255–265.

Cathcard, C., Carling, G., Larson, F., Johansson, R., & Round, E. (2018). Areal pressure in grammatical evolution. An indo-european case study. *Diachronica*, *35*(1), 1–34.

Chén, B. 陈. (1996). *Lùn yǔyán jiēchù yǔ yǔyán liánméng* 论语言接触与语言联盟 [Language contact and language unions]. Běijīng: Yǔwén.

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

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

Dagan, T., & Martin, W. (2007). Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(3), 870–875.

Daval-Markussen, A., & Bakker, P. (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 S. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11)*.

Dellert, J. (2018). Combining information-weighted sequence alignment and sound correspondence models for improved cognate detection. In *Proceedings of the 27th international conference on computational linguistics* (pp. 3123–3133).

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

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://wals.info/X

Eger, S., & Mehle, A. (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* (pp. 52–58). Association for Computational Linguistics.

Fitch, W. M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology*, *20*(4), 406–416.

Friedman, V. A. (2007). Balkanizing the Balkan Sprachbund. In A. Y. Aikhenvald & R. M. W. Dixon (Eds.), *Grammars in contact: A cross-linguistic typology* (Vol. 4, pp. 201–219). Oxford: Oxford University Press.

Galucio, A. V., Meira, S., Birchall, J., Moore, D., Gabas Júnior, N., Drude, S., … Rodrigues, C. R. (2015). Genealogical relations and lexical distances within the Tupian linguistic family. *Boletim Do Museu Paraense Emílio Goeldi. Ciências Humanas*, *10*, 229–274.

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

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

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

Hantgan, A., & List, J.-M. (n.d.). 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.

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance* (PhD thesis). Rijksuniversiteit Groningen; Rijksuniversiteit Groningen, Groningen.

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

Holman, E. W., Brown, C. H., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., … Egorov, D. (2011). Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, *52*(6), 841–875. Retrieved from http://www.jstor.org/stable/10.1086/662127X

Huson, D. H. (1998). SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics*, *14*(1), 68–73.

Jacques, G., & List, J.-M. (2019). Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them). *Journal of Historical Linguistics*, 19(1), 128-167.

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 B. Kaplan & S. Wapner (Eds.), *Perspectives in psychological theory: Essays in honor of Heinz Werner* (pp. 124–134). New York: International Universities Press.

Jarceva, V. N. (Ed.). (1990). *Lingvističeskij ėnciklopedičeskij slovar (Linguistical encyclopedical dictionary)*. Moscow: Sovetskaja Enciklopedija.

Kaiping, G. A., & Klamer, M. (2018). LexiRumah: An online lexical database of the lesser sunda islands. *PLOS ONE*, *13*(10), 1–29. doi:10.1371/journal.pone.0205250X

Kluge, F. (Ed.). (2002). *Etymologisches Wörterbuch der deutschen Sprache* (24th ed.). Berlin: CD-ROM; de Gruyter.

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* (pp. 288–295).

Kondrak, G. (2002). Determining Recurrent Sound Correspondences by Inducing Translation Models. In *Nineteenth International Conference on Computational Linguistics (COLING 2002)* (pp. 488–494). Taipei. Retrieved from http://www.cs.ualberta.ca/~kondrak/papers/cic03.pdfX

Lass, R. (1997). *Historical linguistics and language change*. Cambridge: Cambridge University Press.

Lee, Y.-J., & Sagart, L. (2008). No limits to borrowing: The case of Bai and Chinese. *Diachronica*, *25*(3), 357–385.

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* (pp. 117–125). Stroudsburg.

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

List, J.-M. (2015). Network perspectives on Chinese dialect history. *Bulletin of Chinese Linguistics*, *8*, 42–67.

List, J.-M. (2016). Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*, *1*(2), 119–136. doi:10.1093/jole/lzw006X

List, J.-M. (2019). Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics*, *1*(45), 1–24. doi:10.1162/coli_a_00344X

List, J.-M., Cysouw, M., & Forkel, R. (2016). Concepticon. A resource for the linking of concept lists. In N. C. (Conference Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 2393–2400). European Language Resources Association (ELRA).

List, J.-M., Greenhill, S. J., Anderson, C., Mayer, T., Tresoldi, T., & Forkel, R. (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.

List, J.-M., Greenhill, S. J., & Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLOS ONE*, *12*(1), 1–18.

List, J.-M., Nelson-Sathi, S., Geisler, H., & Martin, W. (2014). Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *Bioessays*, *36*(2), 141–150.

List, J.-M., Nelson-Sathi, S., Martin, W., & Geisler, H. (2014). Using phylogenetic networks to model Chinese dialect history. *Language Dynamics and Change*, *4*(2), 222–252.

List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T., & Forkel, R. (2018). Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, *3*(2), 130–144.

Liú, Lìlǐ 刘俐李, Wáng, Hóngzhōng 王洪钟, & Bǎi, Yíng 柏莹. (2007). *Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí*. Nánjīng: Fènghuáng.

Maddieson, I., Flavier, S., Marsico, E., Coupé, C., & Pellegrino., F. (2013). LAPSyD: Lyon-Albuquerque Phonological Systems Database. In *Proceedings of Interspeech*.

Makarenkov, V., Boc, A., Delwiche, C. F., Diallo, A. B., & Philippe, H. (2006). New efficient algorithm for modeling partial and complete gene transfer scenarios. In V. Batagelj, H.-H. Bock, A. Ferligoj, & A. Žiberna (Eds.), *Data science and classification* (pp. 341–349). Berlin; Heidelberg: Springer Berlin Heidelberg.

Malkiel, Y. (1954). Etymology and the Structure of Word Families. *Word*, *10*(2-3), 265–274.

Matisoff, J. A. (2015). *The Sino-Tibetan Etymological Dictionary and Thesaurus project*. Berkeley: University of California.

McMahon, A., Heggarty, P., McMahon, R., & Maguire, W. (2007). The sound patterns of Englishes: Representing phonetic similarity. *English Language and Linguistics*, *11*(1), 113–142.

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

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

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

Meillet, A. (1925). *Linguistique historique et linguistique générale*. Paris: Libr. Champion.

Mennecier, P., Nerbonne, J., Heyer, E., & Manni, F. (2016). A Central Asian language survey. *Language Dynamics and Change*, *6*(1), 57–98.

Minett, J. W., & Wang, W. S.-Y. (2003). On detecting borrowing. *Diachronica*, *20*(2), 289–330.

Moran, S., McCloy, D., & Wright, R. (2014). PHOIBLE Online. Leipzig: Max Planck Institute for Evolutionary

Anthropology.

Morrison, D. A. (2014). Phylogenetic networks: A new form of multivariate data summary for data mining and exploratory data analysis. *WIREs Data Mining and Knowledge Discovery*. doi:doi: 10.1002/widm.1130X

Nakhleh, L. (2013). Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology & Evolution*, 28(12). 719-728.

Nakhleh, L., Ringe, D., & Warnow, T. (2005). Perfect Phylogenetic Networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, *81*(2), 382–420.

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

Nelson-Sathi, S., List, J.-M., Geisler, H., Fangerau, H., Gray, R. D., Martin, W., & Dagan, T. (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., Colen, R., Gooskens, C., Kleiweg, P., & Leinonen, T. (2011). Gabmap – A web application for dialectology. *Dialectologia*, *Special Issue II*, 65–89.

Nichols, J. (2003). Diversity and stability in language. In Joseph, B. D. and Janda, R. D. (eds). *The handbook of historical linguistics*. Malden: Blackwell (pp. 283-310).

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

Pfeifer, W. (Ed.). (1993). *Etymologisches Wörterbuch des Deutschen* (2nd ed., Vols. 1-2). Berlin: Akademie.

Polyakov, V. N., & Solovyev, V. D. (2006). *Kompjuternye modeli i metody v tipologii i komparativistike*. Kazan': Kazan'skij universitet.

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

Prokić, J., Wieling, M., & Nerbonne, J. (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* (pp. 18–25).

Schmidt, J. (1872). *Die Verwantschaftsverhältnisse der indogermanischen Sprachen* [On the genetic relations among the Indo-European languages]. Leipzig: Hermann Böhlau.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *1*, 195–197.

Starostin, G. S. (Ed.). (2008). Tower of Babel: An etymological database project. Online ressource.

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

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

Steiner, L., Stadler, P. F., & Cysouw, M. (2011). A pipeline for computational historical linguistics. *Language Dynamics and Change*, *1*(1), 89–127.

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

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. Retrieved from http://www.jstor.org/stable/3143802X

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, *21*(2), 121–137. Retrieved from http://www.jstor.org/stable/1263939X

Szeto, P. Y., Ansaldo, U., & Matthews, S. (2018). Typological variation across mandarin dialects: An areal perspective with a quantitative approach. *Linguistic Typology*, *22*(2), 233–275. doi:10.1515/lingty-2018-0009X

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. In M. Haspelmath & U. Tadmor (Eds.), *Loanwords in the world's languages: A comparative handbook* (pp. 55–75). Berlin; New York: de Gruyter.

Wang, W. S.-Y. (2006). *Yǔyán, yǔyīn yǔ jìshù*. Shànghǎi: Xiānggǎng Chéngshì Dàxué.

Whewell, W. D. D. (1847). *The philosophy of the inductive sciences, fofound upon their history* (2nd ed., Vol. 2). London: John W. Parker.

Wieling, M., Margaretha, E., & Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, *40*(2), 307–314.

Willems, M., Lord, E., Laforest, L., Labelle, G., Lapointe, F.-J., Di Sciullo, A. M., & Makarenkov, V. (2016). Using hybridization networks to retrace the evolution of Indo-European languages. *BMC Evolutionary Biology*, *16*(1), 1–18. doi:10.1186/s12862-016-0745-6X

Willis, D. (2011). Reconstructing last week's weather: Syntactic reconstruction and brythonic free relatives. *Journal of Linguistics*, *47*(2), 407–446.

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

Zenner, E., Speelman, D., & Geeraerts, D. (2014). Core vocabulary, borrowability and entrenchment. *Diachronica*, *31*(1), 74–105.

Zhang, L., Manni, F., Fabri, R., & Nerbonne, J. (2019). Detecting loan words computationally. Amsterdam: Draft, submitted to the Contact Language Libraries series; Benjamins.

Supplementary Material

The supplementary material consists of a short tutorial that illustrates how some methods for automated borrowing detection can be applied with help of the Python programming languages and available software libraries. It is online available from: https://zenodo.org/record/3236495

Acknowledgments

This research was funded by the the ERC Starting Grant 715618 Computer-Assisted Language Comparison (http://calc.digling.org). I thank the two anonymous reviewers for constructive and helpful comments.

A short tutorial on different techniques for automatic borrowing detection

Installation requirements

You will need LingPy (List et al. 2018), https://github.com/lingpy/lingpy, in its most recent version. You can install LingPy with the package manager `pip` LingPy by typing:

```
$ pip install lingpy
```

Preparing the data

We start by importing all relevant libraries.

```
from lingpy import *
from itertools import combinations
from collections import defaultdict
from lingpy.compare.phylogeny import PhyBo
```

Now, we load the file, and determine the languages we want to use for the study.

```
lex = LexStat('IEL.csv')
languages = ['English', 'German', 'French', 'Spanish', 'Dutch_List',
        'Italian', 'Breton_ST']
lex.output('tsv', filename='iel-subs',
        subset=True,
        rows=dict(doculect='in '+str(languages)))
```

Computing distances to create data for SplitsTree

Now we can calculate distances, to be shown in the SplitsTree software package.

```
lex = LexStat('iel-subs.tsv')
lex.distances = lex.get_distances(method='sca', aggregate=True)
lex.output('dst', filename='distances')
```

The resulting file can be directly imported in SplitsTree.

Determine cognates and align the data

By identifying automatic cognates, we can also find sequences which are "too similar" to be cognate, like the example for English *person*.

```
lex.cluster(method='sca', threshold=0.45, ref='scaid')
alms = Alignments(lex, ref='scaid')
alms.align()
alms.output('tsv', filename='aligned')
```

To open and search the file, we recommend to use the EDICTOR tool (List, Greenhill, and Gray 2017) at http://edictor.digling.org, which provides an easy access to inspect the alignments.

ID	DOCULECT	CONCEPT	SCAID	ALIGNMENT
2448	Breton	person	632	d ẽː n
2450	Dutch	person	633	m ɛ n s
2444	German	person	633	m ɛ n ʃ
2460	Spanish	person	638	o m b r e
2445	Italian	person	634	p e r s o n a
2451	French	person	634	p ɛ ʀ s ɔ n -
2447	English	person	634	p ɜː - s ə n -

Figure 1: Alignment in EDICTOR software for "person".

Computing minimal lateral networks

To compute minimal lateral networks of the data, we can also use the LingPy software. We export one plot of the data, showing the inferred evolution of the words for 'person' in the data.

```
phy = PhyBo('aligned.tsv', ref='cogid', tree_calc='upgma')
phy.analyze()
phy.plot_concept_evolution('w-2-1', 'person', radius=0.8, outer_radius=0.1,
        proto='alignment', fileformat='pdf')
```

Here is the resulting plot:

References

List, Johann-Mattis, Simon J. Greenhill, and Russell D. Gray. 2017. "The Potential of Automatic Word Comparison for Historical Linguistics." *PLOS ONE* 12 (1): 1–18.

List, Johann-Mattis, Simon Greenhill, Tiago Tresoldi, and Robert Forkel. 2018. "LingPy. A Python Library for Quantitative Tasks in Historical Linguistics." Jena: Max Planck Institute for the Science of Human History. 2018. http://lingpy.org.

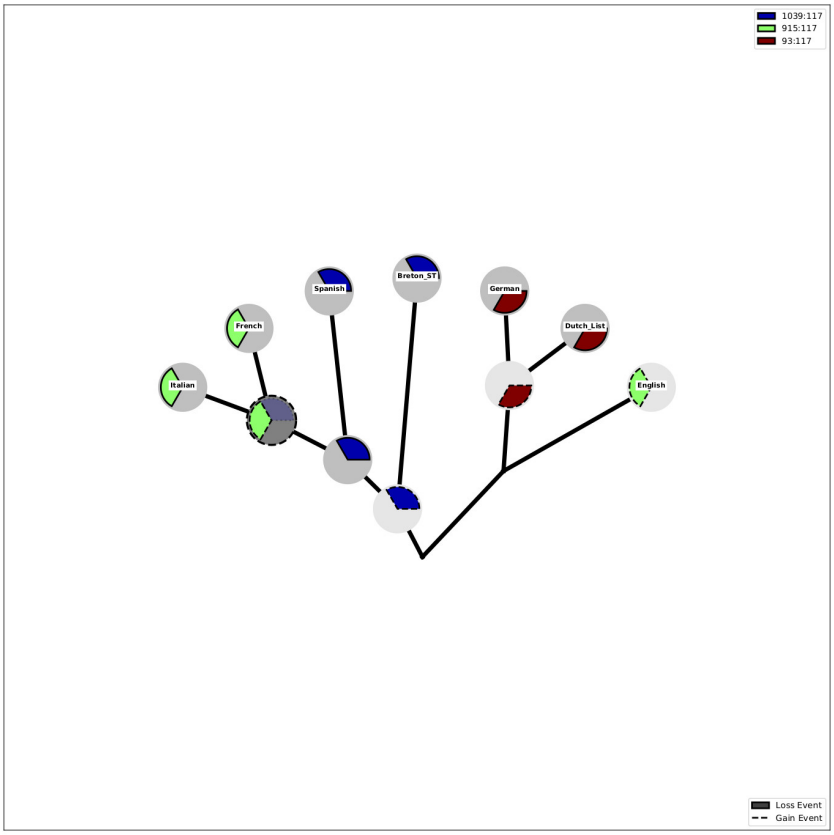


Figure 2: Minimal lateral network for “person”.