Check for updates

RESEARCH ARTICLE

REVISED Automated identification of borrowings in multilingual wordlists [version 3; peer review: 4 approved]

Johann-Mattis List, Robert Forkel

Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Thüringen, 04103, Germany

v3 **First published:** 15 Jul 2021, **1**:79
https://doi.org/10.12688/openreseurope.13843.1

Second version: 24 Aug 2021, **1**:79
https://doi.org/10.12688/openreseurope.13843.2

Latest published: 23 Mar 2022, **1**:79
https://doi.org/10.12688/openreseurope.13843.3

Abstract

Although lexical borrowing is an important aspect of language evolution, there have been few attempts to automate the identification of borrowings in lexical datasets. Moreover, none of the solutions which have been proposed so far identify borrowings across multiple languages. This study proposes a new method for the task and tests it on a newly compiled large comparative dataset of 48 South-East Asian languages from Southern China. The method yields very promising results, while it is conceptually straightforward and easy to apply. This makes the approach a perfect candidate for computer-assisted exploratory studies on lexical borrowing in contact areas.

Keywords

computational linguistics, historical linguistics, lexical borrowing, borrowing detection, computational historical linguistics

This article is included in the Languages and Literature gateway.

This article is included in the European Research Council (ERC) gateway.

Open Peer Review

Approval Status ✓ ✓ ✓ ✓

	1	2	3	4
version 3 (revision) 23 Mar 2022				
version 2 (revision) 24 Aug 2021		✓ view ↑		
version 1 15 Jul 2021	✓ view	? view	✓ view	✓ view

1. **Pui Yiu Szeto**, The University of Hong Kong, Hong Kong, China
2. **John Nerbonne**, University of Groningen, Groningen, The Netherlands
3. **George S. Starostin**, Russian State University for the Humanities, Moscow, Russian Federation
4. **Kenichi W. Okamoto**, University of St. Thomas, St. Paul, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Linguistic Diversity collection.

Corresponding author: Johann-Mattis List (mattis_list@eva.mpg.de)

Author roles: List JM: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Forkel R**: Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715618)
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: List JM and Forkel R. **Automated identification of borrowings in multilingual wordlists [version 3; peer review: 4 approved]** Open Research Europe 2022, **1**:79 https://doi.org/10.12688/openreseurope.13843.3

First published: 15 Jul 2021, **1**:79 https://doi.org/10.12688/openreseurope.13843.1

REVISED **Amendments from Version 2**

In this version, we have tried to account for the remarks by reviewer John Nerbonne, adding a small paragraph towards the end of the General Results section, where we speculate why our approach performs better than the alternative approaches.

Any further responses from the reviewers can be found at the end of the article

Plain language summary

Lexical borrowing, the transfer of words from one language to another, is one of the most prominent aspects of language evolution. Despite its prominence, only a few attempts have been made to create computational methods that would identify borrowings automatically from linguistic datasets. In this study, we propose a new method which is straightforward and easy to apply. We test it on a dataset of 48 languages from Southern China and find that it yields good results. We conclude that the method may be useful for future computer-assisted studies on lexical borrowing.

Introduction

Few phenomena in linguistics are as pervasive as language contact (Onysko, 2019). It is the first factor that needs to be excluded when searching for genealogical language relationships or universals in the languages of the world. It is an indispensable aspect of studies on human cognition, since any study trying to explain the human language faculty must explain how humans can master a multitude of languages. Language contact is so widespread that it was the first factor of language change identified by early philosophers (compare Plato's *Kratylos* dialogue), more than two millennia before scholars began to understand that all languages are subject to change even without contact (Schleicher, 1863). Due to its pervasiveness, language contact is also a powerful witness to human prehistory, as illustrated by phonetically similar names for the sweet potato in Polynesian and Quechuan languages, which provide evidence that its transfer to Polynesia was due to human contact (Montenegro *et al.*, 2008).

While comparative linguistics has experienced a quantitative turn during the past decades, studies on language contact are still almost exclusively carried out manually, and quantitative studies of language contact phenomena are still in their infancy (List, 2019a). This also applies to automated methods for the identification of lexical borrowings. Although some methods have been proposed, none of them deal with *multilingual wordlists*. In this study, we propose a new method for this task and test it on a newly compiled dataset of South-East Asian languages. The method yields very promising results, while it is conceptually straightforward and easy to apply. This makes the approach a perfect candidate for computer-assisted exploratory studies on lexical borrowing in contact areas.

Background

Since historical linguists typically try to exclude borrowings from their analysis rather than making borrowings part of their analysis, methods for the identification of language contact situations have never really left the "shortcut status", where scholars make use of a certain number of ad-hoc criteria to discuss the degree of language contact in a certain region. It is therefore not surprising that computational methods for the identification of borrowed traits are still in their infancy (List, 2019a), although computational methods in historical linguistics have been flourishing in the past decades. Of the few methods which have been proposed so far, there are phylogenetic network approaches which do not require a strict tree-like phylogeny to model language evolution, but instead assume that certain traits can also be transferred laterally through contact (List, 2015; List *et al.*, 2014a; List *et al.*, 2014b; Nakhleh *et al.*, 2005; Nelson-Sathi *et al.*, 2011). While most phylogenetic network approaches deal with lexical data and try to infer lexical borrowings, recent studies have shown that these approaches can likewise be used to study the areal spread of grammatical traits (Cathcart *et al.*, 2018).

On the other hand, scholars have tried to identify borrowings directly with techniques for automated sequence comparison. These methods treat phonetically transcribed words in spoken languages as sound sequences and then seek to identify similar sequences by using techniques originally designed for computer science and evolutionary biology (List, 2014). Since sequence comparison techniques are primarily applied to identify cognate words (words shared by common inheritance), most methods that make use of them can only identify borrowings between genetically unrelated languages (van der Ark *et al.*, 2007; Mennecier *et al.*, 2016; Zhang *et al.*, 2021) and only a few attempts have been made to identify borrowings in genetically related languages (Hantgan & List, forthcoming).

Materials and methods

Materials

For this study, a new dataset was compiled by aggregating several existing datasets on South-East Asian languages spoken in Southern China. The core of the dataset is a collection of 25 Hmong-Mien language varieties documented by Chén (2012). This dataset was standardized in an earlier study (Wu *et al.*, 2020) by converting it to the standard formats recommended by the Cross-Lingustic Data Formats initiative (CLDF, Forkel *et al.*, 2018). Using the CLDFBench toolkit (Forkel & List, 2020) allows regular and transparent data conversion to CLDF including links to reference catalogs, such as Glottolog for language varieties (Hammarström *et al.*, 2021) and Concepticon for concepts (List *et al.*, 2021a). In addition, CLDF makes transcriptions transparent by linking segments to the B(road)IPA transcription system, which is a stricter version of the standard International Phonetic Alphabet (IPA, 1999), for the representation of speech sounds (Anderson *et al.*, 2018; List *et al.*, 2021b).

Having shown earlier that CLDF greatly facilitates the aggregation of data from diverse sources (List *et al.*, 2018), we assembled data on additional South-East Asian language varieties from sources which were either already converted to CLDF in earlier works (Běijīng Dàxué, 1964) or prepared specifically for this study (Wang, 2004). Table 1 shows the eight core

Table 1. Sources of the data selected for this study. Note that due to the rather low coverage of data without base list of 250 concepts in some datasets, several varieties were aggregated from two or more sources. This is indicated in the CLDF version of the dataset. For character readings, where rudimentary Concepticon mapping was carried out on a very selective basis, only the number of concepts linked to Concepticon is shown in the Source column, also indicated by an asterisk.

ID	Source	Family	Varieties		Concepts	
			Source	Selected	Source	Selected
beidasinitic	Běijīng Dàxué (1964)	Chinese dialects	18	6	905	146
beidazihui	Běijīng Dàxué (1962)	Chinese dialects (characters)	19	4	*518	171
castrosui	Castro & Pan (2015)	Sui dialects (Tai-Kadai)	16	3	608	211
castroyi	Castro *et al.* (2010)	Loloish dialects (Sino-Tibetan)	6	1	540	222
castrozhuang	Castro & Hansen (2010)	Zhuang dialects (Tai-Kadai)	20	8	511	243
chenhmongmien	Chén (2012)	Hmong-Mien	25	23	888	250
housinitic	Hóu (2004)	Chinese dialects	40	10	180	61
houzihui	Hóu (2004)	Chinese dialects (characters)	40	9	*155	77
liusinitic	Liú *et al.* (2007)	Chinese dialects	19	5	201	130
wangbai	Wang (2004)	Bai dialects (Sino-Tibetan)	9	1	471	144

datasets that were used in this study. In addition, Chinese dialect data is available in the form of lists of character pronunciations. While these do not provide any strict information on actual words, since individual characters correspond to morphemes in Chinese, it can still be useful to include them in larger multilingual collections, since they may fill gaps where the available data on words in Chinese dialects is sparse. For this reason, the character readings from two datasets (Běijīng Dàxué, 1962; Hóu, 2004) were roughly linked to common concepts in our base datasets in order to increase the coverage for individual language varieties.

From these datasets, a subset of 48 language varieties (23 Hmong-Mien languages, 11 Tai-Kadai languages, and 14 Sino-Tibetan languages) and 250 concepts were selected. The criterion for data and concept selection was the general coverage of individual language varieties with respect to the 250 concepts chosen, and the geographic proximity of the languages in the sample. While the coverage with respect to the concepts for which there is a word form in individual language varieties is reasonably high in most cases, there are some outliers, mostly from Chinese dialects, in which more than 40% of the concepts are missing. While a low coverage for certain varieties would be problematic for phylogenetic studies (Sagart *et al.*, 2019), it was nevertheless decided to keep these language varieties in the sample, mostly because their geographic position makes them interesting candidates for donor and recipient languages in the current sample. Figure 1 shows the geographic distribution of the languages in our sample.

While it would be desirable to extend the data further, we consider the current collection as sufficient for initial experiments on automated borrowing detection. Extending the dialect data, specifically for the Chinese dialects, could in theory be achieved in the future by integrating information on individual dialects extracted from dialect dictionaries, such as the collection of Lǐ (2002), which comprises dialect dictionaries for 42 dialect varieties. The integration of this and similar sources, however, cannot be done in a straightforward way, since the data was usually not collected from established questionnaires, but is instead listed only by dialect headwords (given by Chinese characters with their pronunciations), which are furthermore only sporadically translated (see Kurpaska, 2010: 128–183 on the structure of the dictionary collection by Lǐ, 2002). As a result, finding the counterpart for a concept like “sun” requires one to search actively for a word that is potentially cognate with “sun” in the target dialect in a first step, and to verify this entry in a second step with the translation provided in the resource. Thus, although it would be desirable to improve on our current data basis, this endeavour would largely exceed the scope of the current study.

Methods

Automated borrowing detection. The new method for the identification of borrowings in multilingual wordlists is based on a very straightforward workflow that proceeds in two steps. In the first step, traditional methods for cognate detection are used to identify *language-family-internal cognate sets* in the data. In the second step, all language-internal cognate sets are compared across language families and clustered into sets of *potentially borrowed words* once the overall average distance among the cognate sets is below a certain threshold. For the first step, it is useful to use a conservative cognate detection method which searches for deep genealogical similarities among

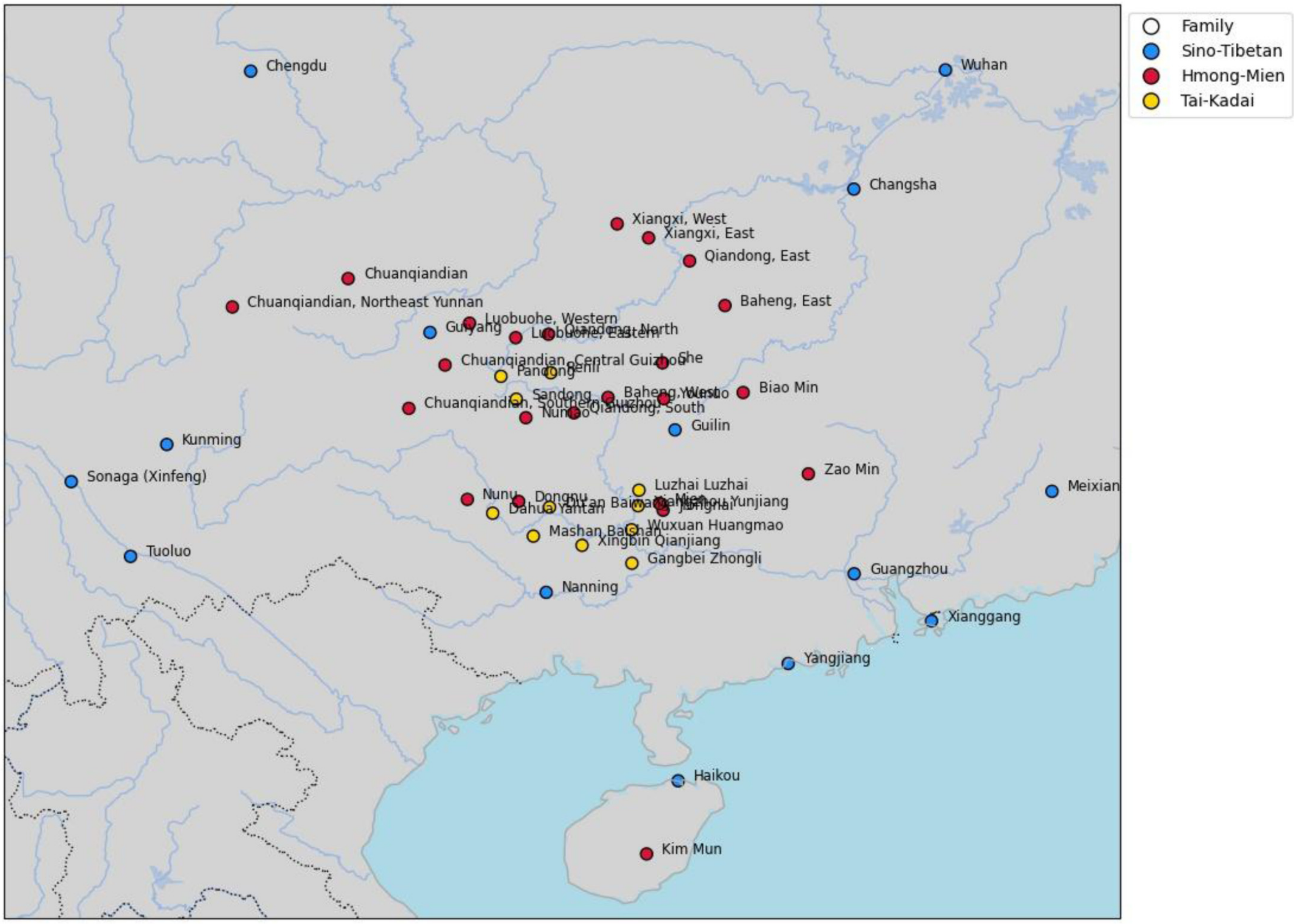


Figure 1. **Languages in the sample.**

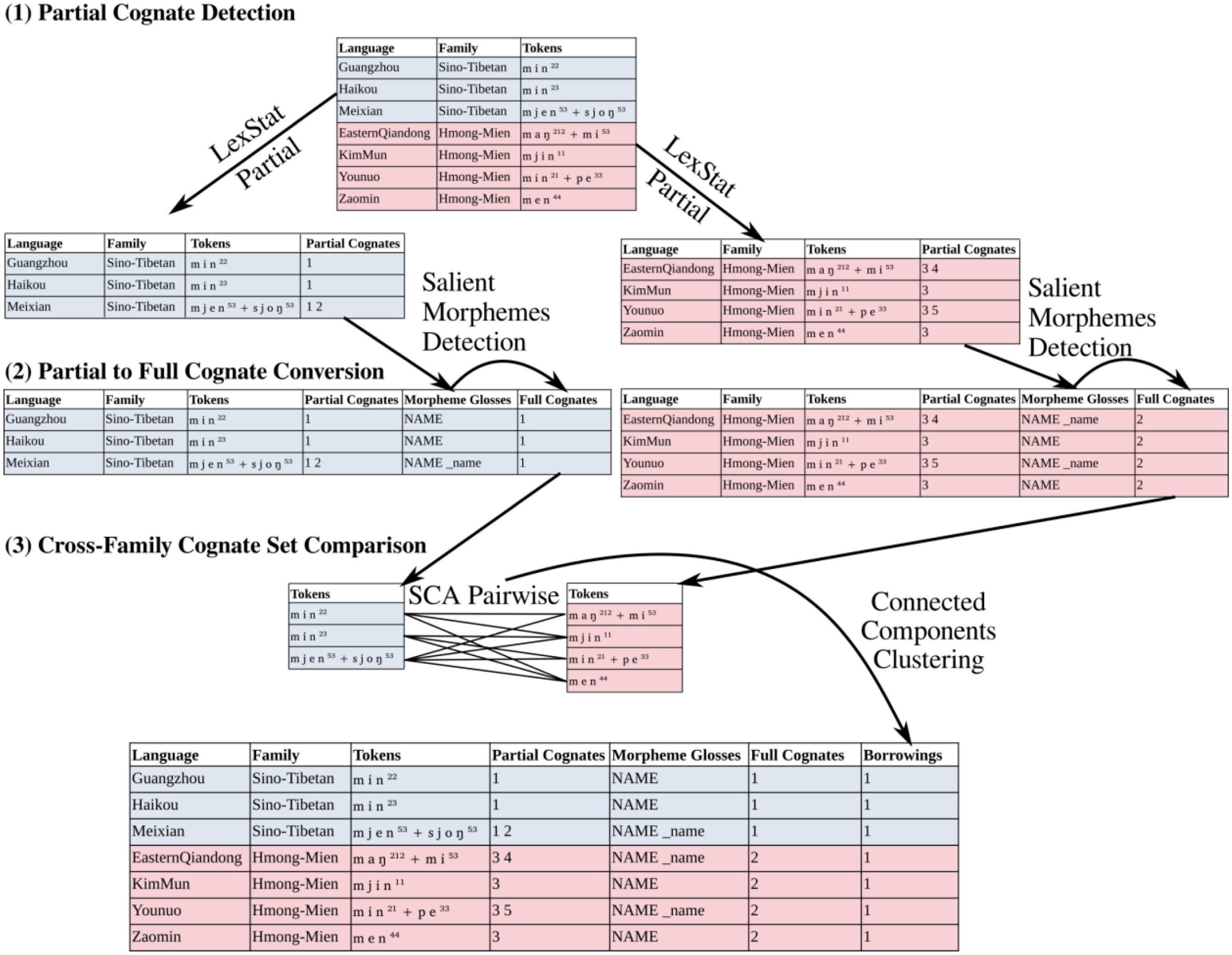
word forms. An example for such a method is the LexStat algorithm for automated cognate detection, which has been proposed earlier (see List, 2012a and List *et al.*, 2017), or its modification, which searches for partial cognates (words which are not entirely cognate but only in parts of their individual morphemes) instead of full cognates (List *et al.*, 2016). For the second step, it is useful to employ a less conservative method for cognate detection that searches for superficial phonetic similarities rather than deep similarities based on regular sound correspondences. Here, the sound-class-based alignment (SCA) method for pairwise and multiple phonetic alignment (List, 2012b) is a good candidate, specifically also because studies on pairwise borrowing detection have shown that SCA outperforms edit distance in this task (Zhang *et al.*, 2021).

For our specific use case, we decided to use the LexStat algorithm adjusted for partial cognate detection (List *et al.*, 2016), since it is well known that South-East Asian languages show frequent compounding patterns which cannot be captured when searching for full cognates in the data (for a detailed discussion on partial cognacy, we refer interested readers to Hill & List, 2017). The LexStat Partial algorithm expects words to be segmented into morphemes by the user and then uses a network approach to cluster morphemes which ocurr in the same concept slot of a given wordlist into sets of cognate morphemes, corresponding to partially cognate words. Since one would expect, however, that borrowings involve full words, it seems useful to employ a full word alignment algorithm for the second stage. When searching for partial cognates in a first instance, this means that one needs to find a way to convert partial cognates to full cognates later. Since words can share cognate morphemes in different parts, the partial cognate relation is not transitive (word *A* with form *xyz* can be partially cognate to word *B* with form *x* and to word *C* with form *ya*, while word *B* is not partially cognate with word *C*, but a word D with form *ab* would be cognate with *D*), a specific conversion procedure that transitivizes cognate relations is needed. For this purpose, we decided to use a new method that we recently developed (Wu & List, 2021). This method is based on a greedy algorithm that assigns partial cognates to full cognate sets which have at least one cognate set in common. This method starts from the most frequently recurring partial cognate, assigning all words which contain this morpheme to the same cognate sets, and then proceeds with the remaining word forms which have not yet been assigned to a cognate set until all words have been visited (yielding the merger of *A* with *B* due to their shared cognate set *x* and then proceed to cluster *C* and *D* due to their shared cognate *a*).

In order to compare two cognate sets expressing the same meaning from two different language families, the new method proposed here first computes pairwise SCA distances for each possible word pair assembled from languages from different language families. The distances are all stored in memory and then averaged. If the average distance is lower than a user-defined threshold, a link between the cognate sets is drawn. After all cognate sets from different languages have been compared in this fashion, the method searches for all *connected components* in this cognate set network and assigns all cognate sets appearing in the same connected component to the same set of potentially borrowed words.

This method does not resolve the direction of borrowings. But instead of earlier approaches, which only identify pairs of potentially borrowed words, it allows to cluster words into *xenologs*, that is, sets of words which are not entirely related by common descent, but also by lateral transfer (List, 2016). Furthermore, since all our methods needs as evidence is to find that two words are attested in different language families (following the idea first expressed in van der Ark *et al.*, 2007), the method might even detect borrowings from a third language family, with the original word (or cognate forms thereof) not being reflected in the sample.

Figure 2 shows a rudimentary workflow example in which the major steps are displayed. In step (1), partial cognates are inferred with the help of the LexStat-Partial algorithm, for each language family in separation. In step (2) the greedy algorithm by Wu & List (2021) is used to convert partial to full cognates



(1) Partial Cognate Detection
LexStat Partial
LexStat Partial
Salient Morphemes Detection
(2) Partial to Full Cognate Conversion
Salient Morphemes Detection
(3) Cross-Family Cognate Set Comparison
SCA Pairwise
Connected Components Clustering

Language	Family	Tokens
Guangzhou	Sino-Tibetan	m i n ^{22}
Haikou	Sino-Tibetan	m i n ^{23}
Meixian	Sino-Tibetan	m j e n ^{53} + s j o ŋ ^{53}
EasternQiandong	Hmong-Mien	m a ŋ ^{212} + m i ^{53}
KimMun	Hmong-Mien	m j i n ^{11}
Younuo	Hmong-Mien	m i n ^{21} + p e ^{33}
Zaomin	Hmong-Mien	m e n ^{44}

Language	Family	Tokens	Partial Cognates
Guangzhou	Sino-Tibetan	m i n ^{22}	1
Haikou	Sino-Tibetan	m i n ^{23}	1
Meixian	Sino-Tibetan	m j e n ^{53} + s j o ŋ ^{53}	1 2

Language	Family	Tokens	Partial Cognates
EasternQiandong	Hmong-Mien	m a ŋ ^{212} + m i ^{53}	3 4
KimMun	Hmong-Mien	m j i n ^{11}	3
Younuo	Hmong-Mien	m i n ^{21} + p e ^{33}	3 5
Zaomin	Hmong-Mien	m e n ^{44}	3

Language	Family	Tokens	Partial Cognates	Morpheme Glosses	Full Cognates
Guangzhou	Sino-Tibetan	m i n ^{22}	1	NAME	1
Haikou	Sino-Tibetan	m i n ^{23}	1	NAME	1
Meixian	Sino-Tibetan	m j e n ^{53} + s j o ŋ ^{53}	1 2	NAME _name	1

Language	Family	Tokens	Partial Cognates	Morpheme Glosses	Full Cognates
EasternQiandong	Hmong-Mien	m a ŋ ^{212} + m i ^{53}	3 4	NAME _name	2
KimMun	Hmong-Mien	m j i n ^{11}	3	NAME	2
Younuo	Hmong-Mien	m i n ^{21} + p e ^{33}	3 5	NAME _name	2
Zaomin	Hmong-Mien	m e n ^{44}	3	NAME	2

Tokens
m i n ^{22}
m i n ^{23}
m j e n ^{53} + s j o ŋ ^{53}

Tokens
m a ŋ ^{212} + m i ^{53}
m j i n ^{11}
m i n ^{21} + p e ^{33}
m e n ^{44}

Language	Family	Tokens	Partial Cognates	Morpheme Glosses	Full Cognates	Borrowings
Guangzhou	Sino-Tibetan	m i n ^{22}	1	NAME	1	1
Haikou	Sino-Tibetan	m i n ^{23}	1	NAME	1	1
Meixian	Sino-Tibetan	m j e n ^{53} + s j o ŋ ^{53}	1 2	NAME _name	1	1
EasternQiandong	Hmong-Mien	m a ŋ ^{212} + m i ^{53}	3 4	NAME _name	2	1
KimMun	Hmong-Mien	m j i n ^{11}	3	NAME	2	1
Younuo	Hmong-Mien	m i n ^{21} + p e ^{33}	3 5	NAME _name	2	1
Zaomin	Hmong-Mien	m e n ^{44}	3	NAME	2	1

Figure 2. Workflow example.

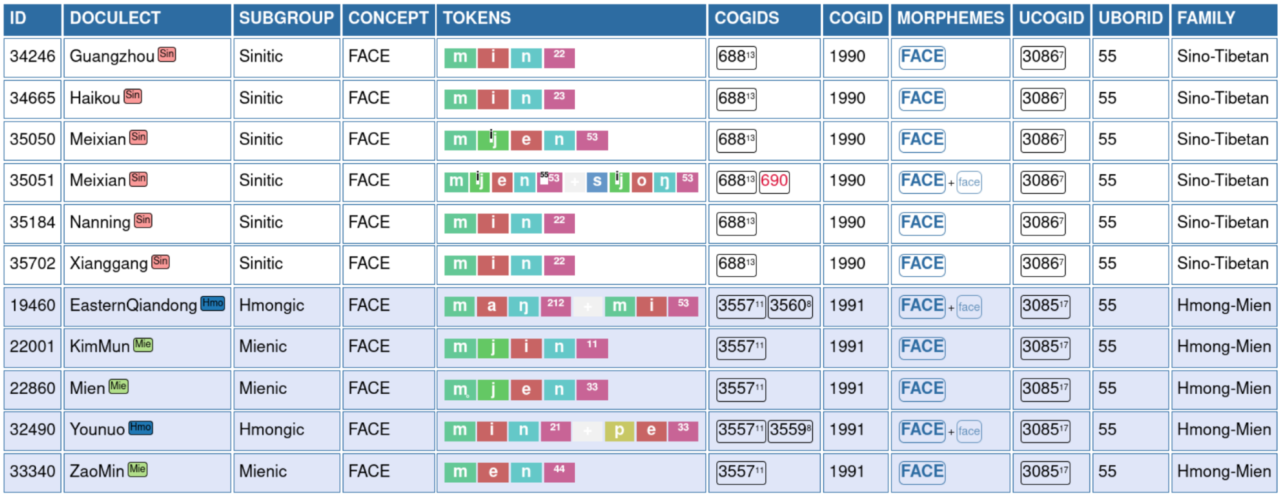
based on the identification of those cognate morphemes which recur in the largest part of the data. In step (3), individual cognate sets across language families are compared with each other using the SCA algorithm for pairwise alignments. Based on these scores (not shown in the figure), a network of cognate sets is constructed in which cognate sets are connected which show an average distance score beyond a user-defined threshold.

The selection of thresholds is important for the identification of partial cognates and for the identification of cross-language-family cognate sets (or sets of xenologous words). Since both thresholds are crucially intertwined with each other, it is not easy to come up with an objective solution for threshold detection. Increasing the threshold for cognate detection, for example, will cluster more words into the same cognate set in related languages. The larger the language-internal cognate sets, however, the higher the chance that borrowings are rejected if the second threshold is considerably low. Since threshold detection in cognate detection tasks is a problem that has no straightforward solution by now, all we can do for now is to derive thresholds by means of trial and error. While this may not seem optimal, we emphasize that we consider the major value of our approach not in the detection of indisputible truths regarding cognate and borrowed words, but rather as a tool supposed to aid linguists in their research, which would necessarily involve a manual refinement of automatic analyses.

An alternative way to achieve results similar to our workflow would be to use the well-established cognate detection algorithms on the whole dataset without splitting the data into language families first, and then infer potential borrowings by identifying those cognate sets which appear in more than one language family. While this approach has been used successfully in a computer-assisted study in the past (Hantgan *et al.*, 2022), our approach has several advantages, as it allows us not only to combine different methods for language-family-internal and language-family-external cognate detection, but also runs faster, since a much smaller part of the data has to be compared in a one-to-one fashion. Since the implementation of this alternative method for borrowing detection can be done in a straightforward way with the help of the LingPy library, we added the code to the code package accompanying this study.

Annotation of borrowings in multilingual wordlists. In order to allow us to test the new method against human judgments, the data was annotated manually, using the EDICTOR (List, 2017), a web-based tool for curation of etymological data in historical linguistics. The data was annotated in two stages. First, cognate sets were identified inside all language families in our sample. In a second stage, cognate sets were themselves assembled into larger sets of potentially borrowed words.

An example for this annotation is shown in Figure 3, where words for "face" in Hmong-Mien and Sino-Tibetan are compared with each other. The table shows our annotation procedure and how data are displayed in the EDICTOR application The ID is used to refer to the original data point and allows us to trace the data from the original sources and across different files. The DOCULECT column provides a language identifier, which also provides rudimentary subgroup information (also displayed in the column SUBGROUP), which was made available with the most recent version of EDICTOR. The column TOKENS shows the sound sequences in segmented, normalized form, segmenting the individual sounds (which may be transcribed by more than one characters) by a space and uses the symbol + as an additional character to indicate morpheme boundaries. Partial cognate identifiers are manually added to the data in the column COGIDS, which are provided on a language-family-internal basis only and indicate with the help of unique identifiers which morphemes of a word form a group of xenologous words. The column MORPHEMES indicates which parts of the word express the main (or salient aspect of the) meaning "FACE" and help us to convert the partial cognates into full cognates, following Wu & List (2021) in taking the salient (or main) morphemes as the criterion to convert partial to



ID	DOCULECT	SUBGROUP	CONCEPT	TOKENS	COGIDS	COGID	MORPHEMES	UCOGID	UBORID	FAMILY
34246	Guangzhou Sin	Sinitic	FACE	m i n 22	688^{13}	1990	FACE	3086^{7}	55	Sino-Tibetan
34665	Haikou Sin	Sinitic	FACE	m i n 23	688^{13}	1990	FACE	3086^{7}	55	Sino-Tibetan
35050	Meixian Sin	Sinitic	FACE	m j e n 53	688^{13}	1990	FACE	3086^{7}	55	Sino-Tibetan
35051	Meixian Sin	Sinitic	FACE	m j e n 55 53 + s j o ŋ 53	688^{13} 690	1990	FACE + face	3086^{7}	55	Sino-Tibetan
35184	Nanning Sin	Sinitic	FACE	m i n 22	688^{13}	1990	FACE	3086^{7}	55	Sino-Tibetan
35702	Xianggang Sin	Sinitic	FACE	m i n 22	688^{13}	1990	FACE	3086^{7}	55	Sino-Tibetan
19460	EasternQiandong Hmo	Hmongic	FACE	m a ŋ 212 + m i 53	3557^{11} 3560^{8}	1991	FACE + face	3085^{17}	55	Hmong-Mien
22001	KimMun Mie	Mienic	FACE	m j i n 11	3557^{11}	1991	FACE	3085^{17}	55	Hmong-Mien
22860	Mien Mie	Mienic	FACE	m̥ j e n 33	3557^{11}	1991	FACE	3085^{17}	55	Hmong-Mien
32490	Younuo Hmo	Hmongic	FACE	m i n 21 + p e 33	3557^{11} 3559^{8}	1991	FACE + face	3085^{17}	55	Hmong-Mien
33340	ZaoMin Mie	Mienic	FACE	m e n 44	3557^{11}	1991	FACE	3085^{17}	55	Hmong-Mien

Figure 3. EDICTOR annotation of borrowings.

full cognates. Column UBORID provides the user judgment on potentially borrowed words. Where known, the SOURCE of a borrowing was also indicated, as illustrated for the FACE by 面, Middle Chinese *mjienH* (Baxter, 1992). Additionally, information on the language family, to which a language belongs, is displayed in the column FAMILY, which is routinely provided in the CLDF datasets, from which the data was aggregated.

Data Formats and Data Representation. Having been collected from datasets originally provided in the unifying CLDF formats, which standardizes elicitation glosses for concepts via Concepticon (List *et al.*, 2021a), language names via Glottolog (Hammarström *et al.*, 2021), and phonetic transcriptions via CLTS (List *et al.*, 2021b), the aggregated dataset along with our manual and automated analyses is also provided in the form of a CLDF dataset, which is curated on GitHub (https://github.com/lexibank/seabor) and archived with Zenodo (Version 1.0, https://doi.org/10.5281/zenodo.5037101). For a detailed introduction of the tabular CLDF formats, we refer the readers to Forkel *et al.* (2018).

As mentioned in the previous section, the annotation of partial cognates, full cognates, and borrowing candidates is carried out with the help of the EDICTOR tool. The format underlying this tool is based on a rather well-established tabular representation of lexical data in which each row of a table is reserved for one word form in a particular language and columns provide individual data on each word form. The EDICTOR database is curated online and can be directly accessed in read-only form by interested scholars (https://digling.org/links/seabor.html). For more information on the format, we refer interested readers to an initial study by Hill & List (2017), where partial cognacy and morpheme glosses were first discussed in detail, as well as the follow-up studies by Schweikhard & List (2020), where these formats were further extended, and Wu & List (2021), where a detailed discussion of the problem of partial cognacy is given.

Implementation. The new method was implemented as part of the LingRex Python package (List & Forkel, 2021a, Version 1.2). LingRex is an extension of LingPy which offers code that is specifically useful for the detection of sound correspondence patterns (List, 2019b) and the prediction of words which have not been elicited from cognate sets (Bodt & List, 2022). With this study, LingRex was further extended by the new method for the identification of borrowed words. The methods used for the efficient manual annotation of borrowed words were introduced as part of the most recent version (2.0.0) of the EDICTOR tool (List, 2021). Plots made in this study were carried out with the help of CLDFViz (Forkel, 2021), a Python package which facilitates the static and interactive visualization of data provided in CLDF, which was expanded to allow for the specific requirements for this study. The supplementary material accompanying this study offers both the data and the code needed to replicate the experiments which are reported in this study. Users who wish to replicate the study reported here can find all necessary information in the file “workflow.md”, where all steps of the workflow (including the creation of figures) are explained.

Results

General results

In order to test the new method for the identification of borrowings in multilingual datasets, the data was analyzed by using the default settings of the partial cognate detection algorithm (threshold of 0.50 and 10000 iterations in the permutation test) in order to search for language-family-internal cognates. We then ran the cross-language-family cognate detection method in order to identify potential borrowings, using a threshold of 0.3 in this step, since this threshold was the optimal one reported for the test of the SCA method for pairwise borrowing detection reported by Zhang *et al.* (2021).

As a first test of the methods, we compared the manually annotated cognates and borrowings with the cognates and borrowings which the automated method identifies, using the B-cubed scores (Amigó *et al.*, 2009) as a measure to compare automated with manual cognate and borrowing judgments in terms of precision, recall, and F-scores. B-cubed scores are a technique for the comparison of two partions of the same data for similarity, and range between 1 and 0. A precision of 1 means that only items which also belong to the same partition in the gold standard are assigned to the same partition in the test, which is identical with no false positive decisions by the algorithm. A recall of 1 means that all items which were assigned to the same partition in the gold standard are also assigned to the same partition in the test, which is identical with no false negative decisions by the algorithm. Combining both scores by taking their harmonic mean, yields the F-score, which has become a standard way to compare cognate detection methods in historical linguistics by now (see List, 2014: 189 for details on the computation).

Despite the intensive degree of contact among South-East Asian languages spoken in Southern China, the number of sets of xenologous words in our data is still rather small, resulting in sparse partitions in which only a small number of words is assigned to larger clusters. In order to compare how well our algorithms perform in borrowing detection, it is therefore important to test the methods against a base line. Here, we follow an idea first expressed in List (2019c) to contrast the results with a lumper and a splitter as a baseline. The lumper assigns all words to the same cognate set, no matter how similar or how dissimilar they are. The splitter assigns all words to different cognate sets. Both baselines help us to put our individual cognate and borrowing detection results into context.

The results of this first experiment show that the method works sufficiently well, as shown in Table 2, reaching F-scores of 0.88 for the automated cognate detection task, and 0.87 for the automated borrowing detection task. The method outperforms the lumper and splitter baselines. The lumper baseline which proves to be most efficient for the cognate detection task achieves 0.72, which is in strong contrast to the F-scores reached by our automated cognate detection method, and the splitter,

Table 2. Results for the evaluation of the automated workflow compared to the gold standard.

Method	Precision	Recall	F-score
automated cognate detection	0.90	0.87	0.88
automated borrowing detection	0.94	0.81	0.87
lumper baseline for cognate detection	0.57	1.0	0.72
lumper baseline for borrowing detection	0.19	1.0	0.32
splitter baseline for cognate detection	1.0	0.25	0.39
splitter baseline for borrowing detection	1.0	0.66	0.80

Table 3. Results for the evaluation of different settings for approaches using a single threshold.

Method	Threshold	Cognates	Xenologs
SCA-Partial	0.35	0.86	0.73
SCA-Partial	0.15	0.76	0.81
SCA	0.40	0.84	0.74
SCA	0.15	0.74	0.85
LexStat-Partial	0.55	0.89	0.75
LexStat-Partial	0.40	0.83	0.83
LexStat	0.6	0.87	0.81
LexStat	0.5	0.84	0.86

which proves to be best on the borrowing detection, where the majority of partitions are singletons, because words cannot be shown to be in a xenologous relation with any other words, receives F-scores of 0.8. That the difference between the splitter baseline and our method is only 0.07 calls for specific attention, since it shows that the fact that words which cannot be assigned to borrowings are assigned to different cognate sets make up for a large part of the high scores, and we can see that the B-cubed scores should never be taken at face value but always compared to a baseline.

Nevertheless, although these results should be taken with a certain care, since no further test sets are available, and no proper division into test and training data has been carried out, the scores can be considered sufficient to prove that the method proposed here is basically useful when searching for words shared across different language families.

In order to explore whether our approach of using two different methods (one “deep” and one “shallow” approach) to identify cognates on the one hand and borrowings on the other hand, we ran a couple of experiments with varying thresholds using the alternative approach to borrowing detection proposed by Hantgan *et al.* (2022) mentioned above. The approach by Hantgan *et al.* (2022) simply compares all words expressing the same concept in all languages with each other and searches for cognate words, using the SCA method for automated cognate detection (List, 2012a) and later identifies those cognate sets as potentially influenced by borrowing which recur across more than one language family. In contrast to our two-step procedure, this method requires only one threshold. In order to test to which degree the choice of the threshold influences the F-scores achieved when searching for family-internal cognates as opposed to searching for cross-family xenologs, we ran the partial and the full cognate detection method with shallow SCA distances and with deep LexStat distances derived from automatically inferred sound correspondences for consecutive thresholds ranging from 0 to 1 in steps of 0.05.

Table 3 shows the thresholds which yield the best scores for the cognate and the xenolog detection task in the shallow (SCA) variant and the deep LexStat variant for partial and full cognates. As can be seen from these results, the two LexStat variants perform better in the cognate detection task. The partial variants outperform their respective non-partial counterparts. In the xenolog detection task, however, the SCA proper approach performs almost as well as the LexStat proper approach (0.85 vs. 0.86), and the variants which search for full as opposed to partial cognates outperform their respective counterparts.

Figure 4 shows plots of the four analyses, contrasting the B-cubed F-scores achieved for the two different tasks. As can be seen from these plots, all approaches reach the best F-scores for different thresholds, depending on the task and the cognate detection task requires larger thresholds than the borrowing detection tasks in all cases. This shows that our approach of splitting the borrowing detection enterprise into two distinct tasks, one that seeks to detect language-family internal cognates and one which compares these cognate sets to find potential borrowing candidates, is generally justified. Although we know well that equating the LexStat approach with a linguist applying the comparative method to identify cognates would go too far, we still think that one of the reasons why our method improves over the single-threshold approaches discussed above is that it comes close in imitating how linguists would proceed when identifying borrowings across multiple languages from different families: they would identify cognates inside the families first, and they would compare groups of words instead of comparing words on a pairwise basis. Our new workflow accounts for this more closely than any single-threshold approach could do. However, given that the workflow has been only tested on one dataset so far, it is important to emphasize that the last word on this method has not been spoken yet and that new tests will be needed in the future to make sure that it works generally better than alternative approaches (and not only on one dataset).

Specific results

Assuming that the method works well enough to capture at least recent borrowing events in our dataset on South-East

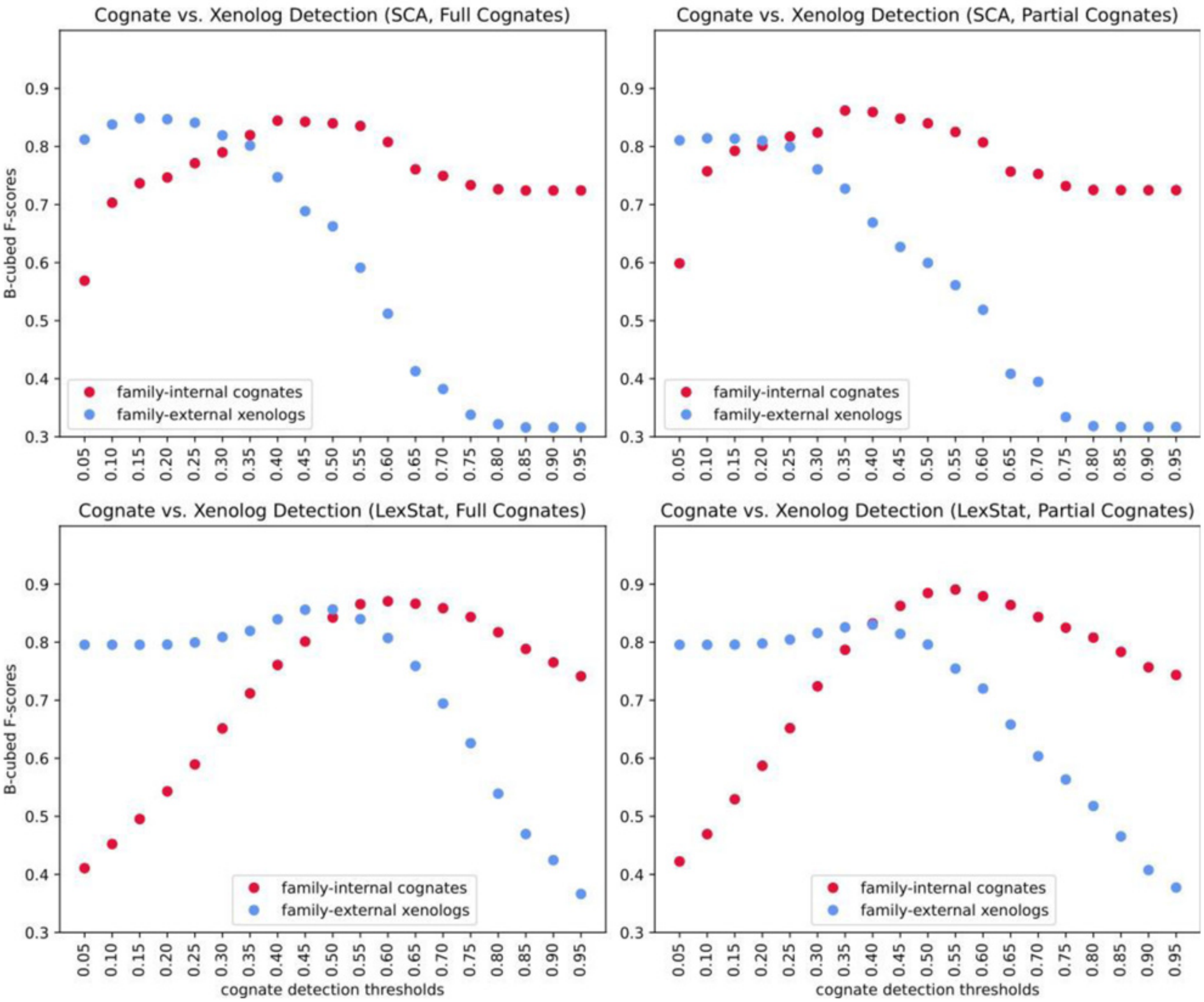


Figure 4. Comparing different settings and thresholds.

Asian languages in Southern China, it is interesting to check whether these borrowings reflect specific patterns. In the linguistic literature it has, for example, for a long time been assumed that certain words are more resistant to borrowing than other words, mostly due to the meanings they express (Swadesh, 1952; Swadesh, 1955). Given that the dataset was assembled from individual CLDF datasets which are themselves linked to the Concepticon project (List *et al.*, 2021a) which in turn provides direct access to a large number of concept lists that have been proposed in the past, it is not difficult to compare to which degree concepts which have been assigned to lists of supposedly stable items behave differently with respect to the automatically inferred borrowings in this sample.

Two supposedly stable concept lists with a high resistance to borrowing are Swadesh's list of 100 items (Swadesh, 1955) and the so-called Leipzig-Jakarta list derived from the World Loanword Database which lists manually identified borrowings for a sample of 41 genetically diverse language varieties (Haspelmath & Tadmor, 2009). In Table 4, we have calculated the average number of non-borrowed items (words which occur uniquely in one language variety and words which are shared within one language family alone) for the traditional Swadesh list of 100 items, the Leipzig-Jakarta list of 100 items (Tadmor, 2009), their respective counterparts (the subset of items which do not occur in the 100-item Swadesh list and the Leipzig-Jakarta list), as well as the base list of 250 concepts. As we can see from the table, there is a considerable difference in terms of supposed stability (or resistance to borrowing) when comparing the supposedly stable, borrowing-resistant sublists with their respective counterparts. While the amount of supposedly borrowed words in this sample may seem to be remarkably high, exceeding 15% for both the sublists and the list of all concepts, it should be kept in mind that our approach does not control for directions and inheritance. Since the source and target words of borrowings are not distinguished, the numbers do not indicate the amount of borrowed words, but the amounts of *xenologs*, that is, sets of etymologically related words which have experienced lateral transfer events in their past (see List, 2016).

In order to test whether the observed differences between the proportions of xenologs are significant, or whether they could have alternatively arisen by chance, we ran 10000 trials in which the concept list was split into two parts, reflecting the proportion of the Swadesh list (with 78 items vs. 172 items) and the Leipzig-Jakarta list (61 items vs. 189 items). In all trials, we tested whether one could observe the same or a higher difference between the amount of non-borrowed items and potential xenologous words. The results suggest that it is not very likely to obtain the differences for the Swadesh list by chance. We obtained similar differences in only 3% of all cases. For the Leipzig-Jakarta list, however, the results were slightly different, and we obtained similar results in 7.1% of all trials. While this number is still low, it would not pass a classical significance test.

Table 4. Proportion of potential borrowings in the data and various sublists.

Concept list	Proportion of non-borrowed items	Number of items
Swadesh (1955)	0.80	78
No Swadesh	0.70	172
Leipzig-Jakarta	0.78	61
No Leipzig-Jakarta	0.72	189
All items	0.73	250

That there are — at times even striking — differences between supposedly stable concepts and concepts more prone to borrowing can also be directly seen when visualizing the characteristics of individual words in each language with the help of geographic plots inspired by "admixture plots" in genetics (Pritchard *et al.*, 2000). In this visual representation, we inspect the words in each language in separation and distinguish (1) *missing data* (no word form for a given concept available), (2) *singletons* (words occur only in this specific variety), and (3) *language-family-internal words* (words that are cognate with words from related languages), from (4) *words shared among two language families* (e.g., Sino-Tibetan vs. Tai-Kadai), and (5) *words shared among all three language families* in our sample. Such a plot is shown in Figure 5 for all 250 concepts in the sample. On the bottom left of the figure, three varieties (Zao Min, Gangbei Zho, and Guangzhou) have been additionally

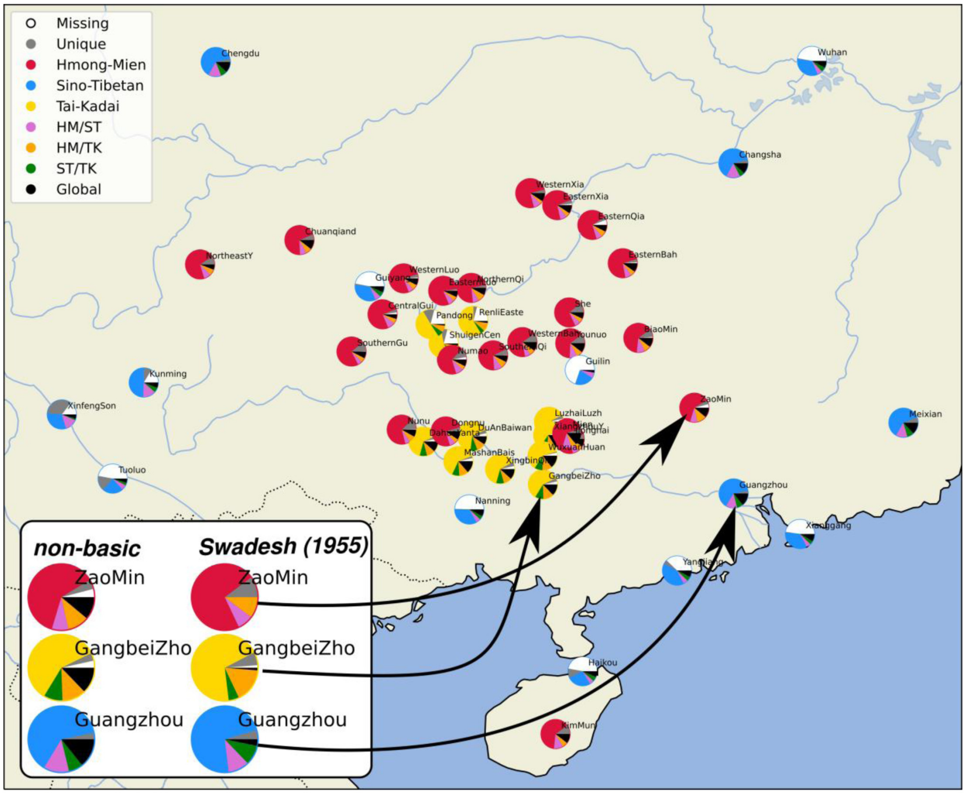


Figure 5. **Admixture plots of shared lexemes between the major language families.**

contrasted with respect to the distribution that they would show for the non-basic items of the 250 concept list and the basic items from Swadesh's list from 1955. As can be seen, all varieties show a much-increased amount of etymologically non-relatable singletons and language-family-internal cognates.

Examples

While numbers and plots can to some degree help us to assess how well a certain method works, it is always important to inspect individual case studies as well in order to explore where the specific weaknesses of a method lie and how this could be overcome in future work. While the EDICTOR interface, introduced earlier, already greatly facilitates the manual inspection and correction of automatically generated cognate judgments (including judgments on potential borrowings), specifically for the detection of borrowings it can be useful to inspect the data in geographic space. For this reason, we created a small routine which plots the inferred sets of words shared across more than one language family on a map and contrasts them with those words which were not assigned to the same cluster.

A first example of this visualization can be seen in Figure 6, showing inferred sets for the concept "name", which are — as we know well — all borrowed from Chinese *míngzì* 名字 (Middle Chinese *mjieng dziH*). While most of the cases inferred by the algorithm are striking (all Tai-Kadai languages have almost literal copies of the Mandarin form), we also find a couple of surprising cases of false negatives in this sample. Thus, the word form [m ei ^{22} + ts ɿ] in the Chinese dialect of Guilin (15 in the center of the map) clearly belongs to the cluster, as does the form [m j ɛ ^{55} + ts ^{h} ɿ] in Xinfeng Sonaga (a Loloish variety of Sino-Tibetan), or the form [m i ^{1}] in Tuluo Bai (a Sino-Tibetan variety whose deeper affiliation remains unclear so far). Since there are two forms in which the Chinese word for "name" can be borrowed, as simplex form *míng* 名, meaning "name", which points to more archaic borrowing events, and in the modern Mandarin form *míngzì* 名字, lit. meaning "name sign", and since the major cluster consists of the bisyllabic form, the algorithm for sequence alignment has problems of identifying short words which lack the final nasal as belonging to the cluster.

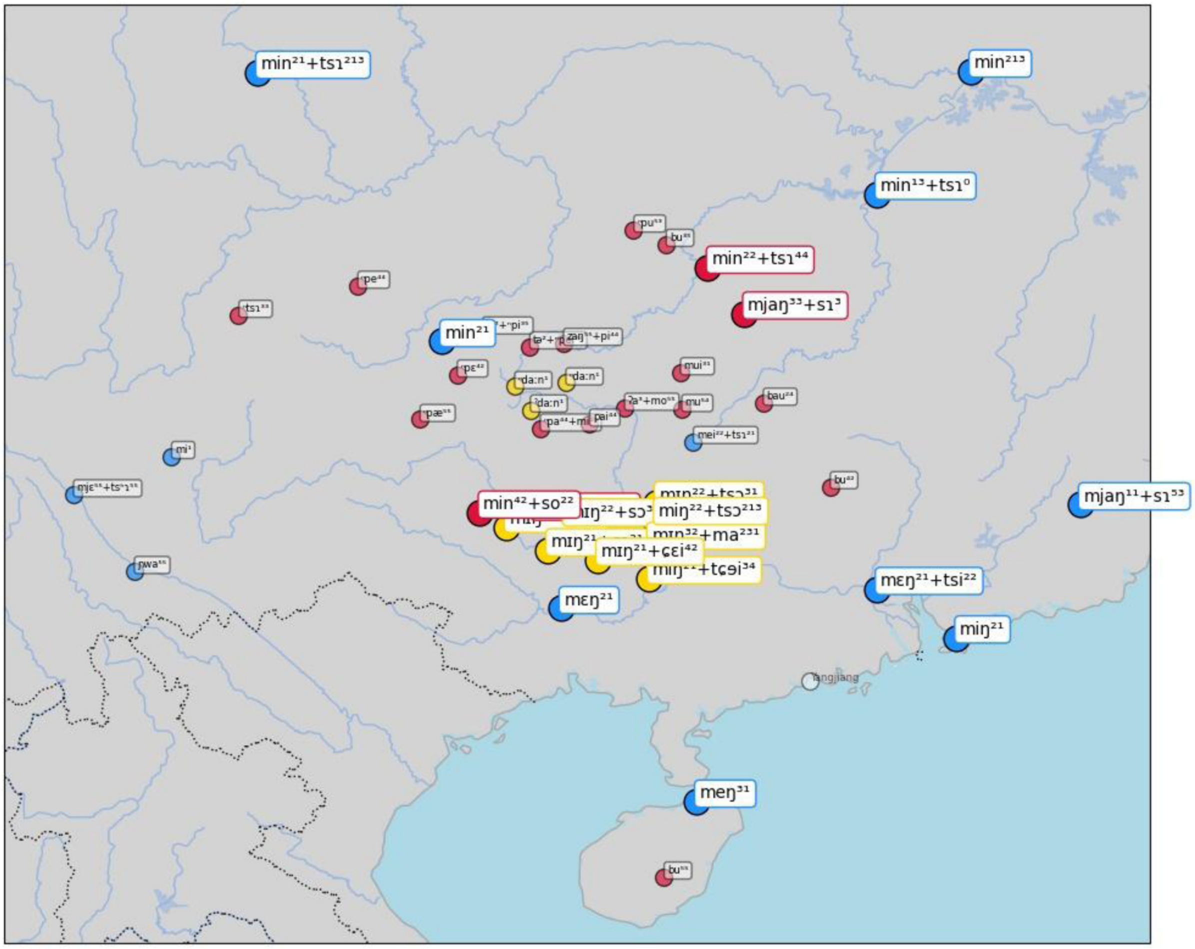


Figure 6. Automatically inferred cluster of potential borrowings for "name".

As a second example, consider words for “flower” in Figure 7. Here, the algorithm correctly identifies the similarity between forms like [w a ^{24}] in the Zhuang varieties of Tai-Kadai, which also occurs in the Hmong-Mien variety Nunu as [v a ^{33}] and is a rather obvious borrowing from Chinese *hua* 花 (Middle Chinese *xwae*), which shows the sound change [xw] > [f] in many Southern varieties of Chinese.

As a last example, consider similar word forms for “correct (right)” in Figure 8. Here the algorithm clusters word forms such as [t ɔi] in BiaoMin (from the Mienic branch of Hmong-Mien) and [t oːi ^{44}] in DahuaYantan (from the Zhuang branch of Tai-Kadai). The source word, however, is again from Chinese, where *duì* 對 (Middle Chinese twoijH) is still the basic way to express “correct (right)” in Mandarin Chinese and many other Chinese dialect varieties. The majority of the sources in the sample selected here provide different words in the Chinese dialects, and we find expressions such as [ts ə n ^{13} + x au ^{53}] (Mandarin Chinese *zhēnhǎo* 真好 “totally right”) in Chengdu. The fact that we find different word forms in the data does not mean, however, that there have been recent events of lexical replacements in many Chinese dialects. It seems instead that this variation is due to the elicitation process. Since the dialect data for Chinese comes from a variety of sources, it is either possible that the mapping of the concepts to the Concepticon project is not entirely correct, or it may be due to the fact that the elicitation process forced the use of longer or more specific expressions, different from the extremely common *duì*.

It would go beyond the scope of this study to discuss all individual findings made by the algorithm in detail. All plots, however, are shared along with the supplementary material accompanying this study, so that interested readers can dive into the individual results and criticize and improve them. What is important to note, however, is that the inspection has shown that there are definite points where the method proposed here could be further improved. The major problem of the scoring procedure used is that it is very sensitive to phonotactic

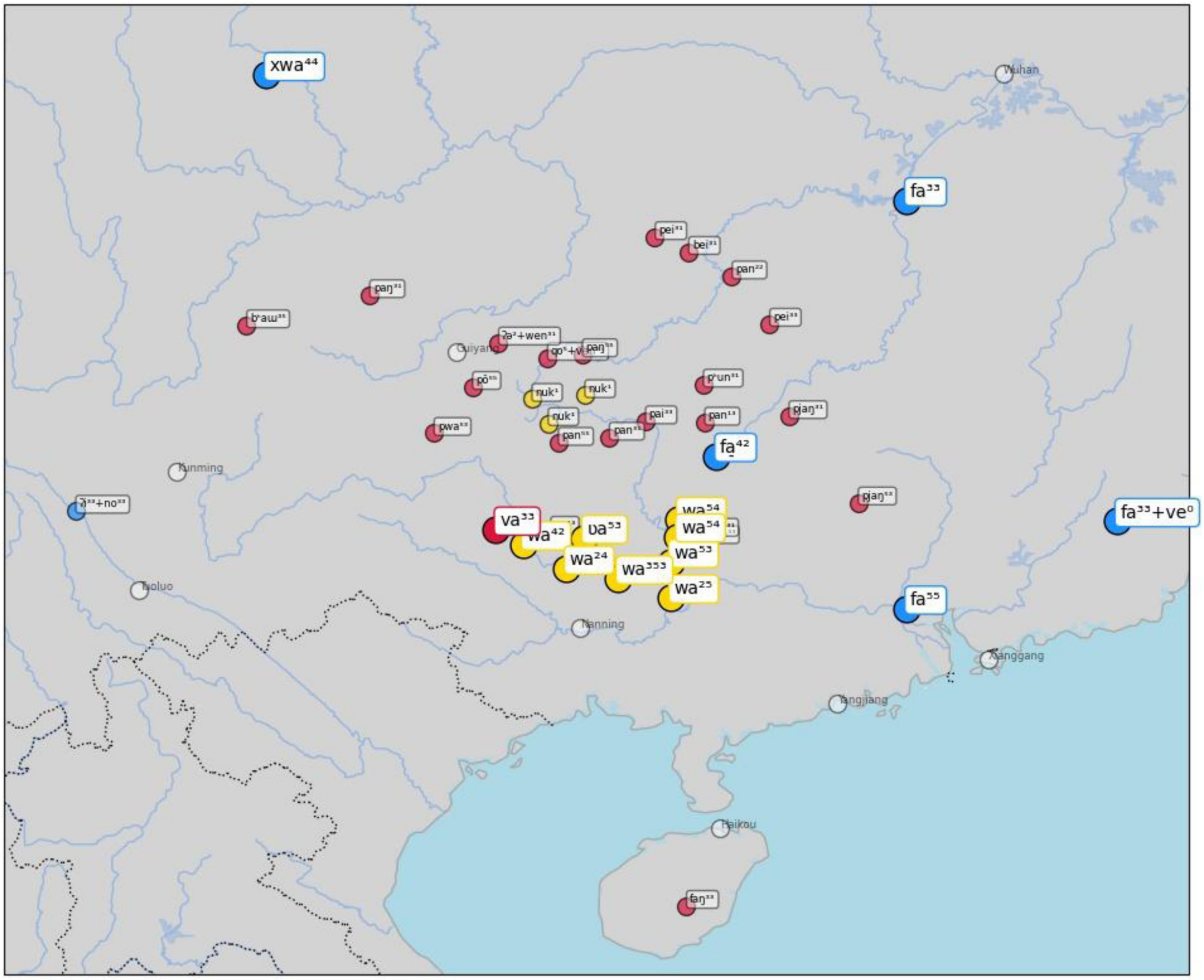


Figure 7. Automatically inferred cluster of potential borrowings for “flower”.

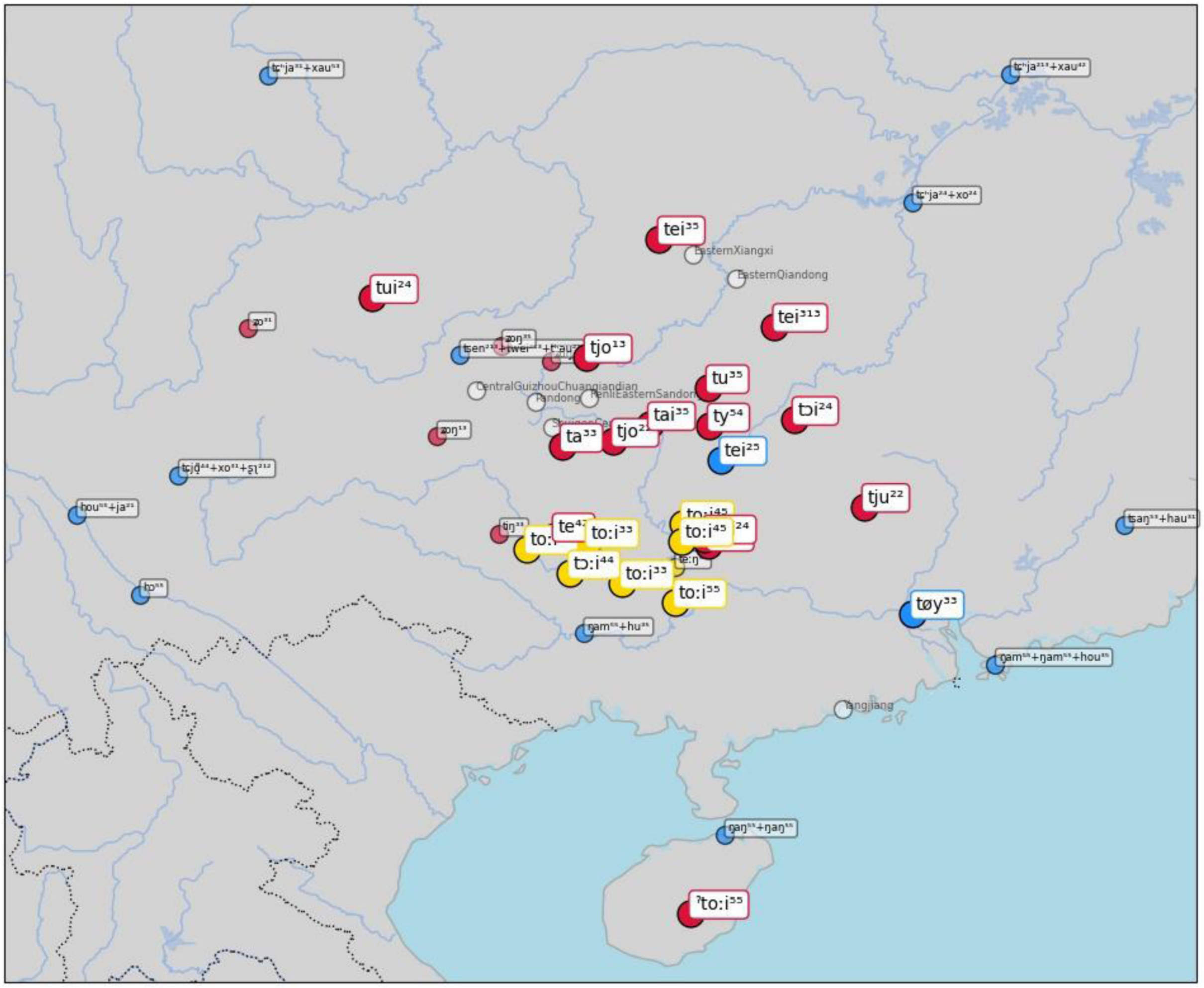


Figure 8. Automatically inferred cluster of potential borrowings for "correct (right)".

representations. As a result, two-word forms which sound rather similar but are represented phonotactically quite different, can be easily judged to be unrelated for the scoring procedure, while human linguists immediately spot the overall similarity. As an example, consider the form [t w əi ^{51}] which is one possible way to represent the word *duì* "correct" in Mandarin Chinese and the form [t oːi ^{55}] in Gangbei Zhongli (Zhuang branch of Tai-Kadai). Phonotactically (for the scoring function), the first form consists of two consonants and one vowel, while the second form only consists of one consonant and one vowel. While humans judging the similarity between both forms would probably ignore the medial [w] as being irrelevant and put more emphasis on the striking match of the diphthongs, the distance scoring employed by the method here is not yet capable of providing such a fine-grained weighting. Alternative approaches for sequence alignments would be needed to detect the particular similarity between the two words, but this is not necessarily trivial, since it would require us to loosen our strict distinction between vowels and consonants, which is crucial in most alignment approaches (including those based on the computation of edit distances, see e.g., Prokić *et al.*, 2009).

Conclusion

Although there has been a lot of research in the field of computational historical linguistics of late, no major improvements in the field of automated borrowing detection have been made so far. The method proposed in this study is very simple and can only detect potential borrowings between languages from different language families. However, as we have tried to show, the method can still be quite useful, both for the automated investigation of contact phenomena in large lexical datasets, and for the more detailed development of computer-assisted case studies, where the method can be used to preprocess the data in order to make the manual annotation more efficient. We therefore consider the new method as a first step towards a more intensive treatment of language contact phenomena in the field of computational historical linguistics.

More detailed case studies will be needed to fully test the potential of our approach. One interesting test case we can think of, could consist of a more detailed study of South-East Asian languages including languages beyond our sample from Southern China, accounting for the observation that specifically Tai-Kadai languages outside of China differ greatly from those inside China, due to the different degrees of Sinitic influence (Szeto & Yurayong, fortcoming). Additional case studies can be carried out with any linguistic area where a sufficient amount of languages from different language families are spoken and lexical data in phonetic transcription for a sufficiently large number of concepts are available.

Data availability

Zenodo: CLDF dataset accompanying List and Forkel's "Borrowing Detection in Multilingual Wordlists" from 2021. https://doi.org/10.5281/zenodo.5037101 (List & Forkel, 2021b)

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Concrete instructions on reproducing the study can be found at https://github.com/lexibank/seabor/blob/v1.0/workflow.md.

Software availability

Source code available from: https://github.com/lexibank/seabor

Archived source code at time of publication: https://doi.org/10.5281/zenodo.5037100 (List & Forkel, 2021b)

License: Creative Commons Attribution 4.0 International

Acknowledgements

We express our gratitude to our four reviewers who all helped a lot to improve the quality of this study.

References

Amigó E, Gonzalo J, Artiles J, *et al.*: **A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints.** *Inf Retrieval.* Hingham, MA USA: Kluwer Academic Publishers. 2009; **12**(4): 461–86.
Publisher Full Text

Anderson C, Tresoldi T, Chacon TC, *et al.*: **A Cross-Linguistic Database of Phonetic Transcription Systems.** *Yearb Pozn Linguist Meet.* 2018; **4**(1): 21–53.
Publisher Full Text

Baxter WH: **A Handbook of Old Chinese Phonology**. Berlin: de Gruyter. 1992.
Publisher Full Text

Běijīng Dàxué: **Hànyǔ Fāngyán Cíhuì** 汉语方言词汇 **[Chinese dialect vocabularies]**. Běijing 北京: Wénzì Gǎigé 文字改革. 1964.

Běijīng Dàxué: **Hànyǔ Fāngyīn Zìhuì** 漢語方音字彙 **[Chinese dialect character pronunciation list]**. Běijīng 北京: Wénzì Gǎigé 文字改革. 1962.

Bodt TA, List JM: **Reflex Prediction. a Case Study of Western Kho-Bwa.** *Diachronica.* 2022; 1–38.
Publisher Full Text

Castro A, Crook B, Flaming R: **A Sociolinguistic Survey of Kua-Nsi and Related Yi Varieties in Heqing County, Yunnan Province, China**. Heqing: SIL International. 2010.
Reference Source

Castro A, Hansen B: **Hongshui He Zhuang Dialect Intelligibility Survey**. Dallas: SIL International. 2010.
Reference Source

Castro A, Pan X: **Sui Dialect Research**. Guizhou: SIL International, 2015.
Reference Source

Cathcart C, Carling G, Larsson F, *et al.*: **Areal Pressure in Grammatical Evolution. An Indo-European Case Study.** *Diachronica.* 2018; **35**(1): 1–34.
Publisher Full Text

Chén Q: **Miàoyáo Yǔwén**. Běijīng: Zhōngyāng Mínzú Dàxué 中央民族大學 [Central Institute of Minorities]. 2012.
Reference Source

Forkel R: **CLDFViz. A python library providing tools to visualize data from CLDF datasets [Computer software, Version 0.5]**. 2021.
Publisher Full Text

Forkel R, List JM: **CLDFBench. Give Your Cross-Linguistic Data a Lift**. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation.* Luxembourg: European Language Resources Association (ELRA). 2020; 6997–7004.
Reference Source

Forkel R, List JM, Greenhill SJ, *et al.*: **Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics.** *Sci Data.* 2018; **5**: 180205.
Publisher Full Text

Hammarström H, Haspelmath M, Forkel R, *et al.*: **Glottolog. Version 4.4**. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Publisher Full Text

Hantgan A, Babiker H, List JML: **First steps towards the detecion of contact layers in Bangime: A multi-disciplinary, computer-assisted approach. [version 1; peer review: awaiting peer review].** *Open Research Europe.* 2022; **2**(10): 1–27.
Publisher Full Text

Hantgan A, List JM: **Bangime: Secret Language, Language Isolate, or Language Island?** Papers in Historical Phonology. forthcoming.
Reference Source

Haspelmath M, Tadmor U: **Loanwords in the World's Languages: A Comparative Handbook**. Berlin; New York: de Gruyter. 2009.
Reference Source

Hill N, List JM: **Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages.** *Yearbook of the Poznań Linguistic Meeting.* 2017; **3**(1): 47–76.
Publisher Full Text

Hóu J: **Xiàndài Hànyǔ Fāngyán Yīnkù** 现代汉语方言音库 **[Phonological database of Chinese dialects].** Shànghǎi 上海: CD-ROM; Shànghǎi Jiàoyù 上海教育. 2004.

IPA: **Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet**. Cambridge: Cambridge University Press. 1999.
Reference Source

Kurpaska M: **Chinese language(s). A look through the prism of The Great Dictionary of Modern Chinese Dialects.** Berlin and New York: De Gruyter. 2010.
Reference Source

Lǐ R: 李荣**. Xiàndài Hànyǔ fāngyán dà cídiǎn** 现代汉语方言大词典 **[The great dictionary of modern Chinese dialects].** Nánjīng: Jiāngsù Jiàoyù. 2002.
Reference Source

List JM: **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*. Stroudsburg. 2012a; 117–25.
Reference Source

List JM: **SCA. Phonetic alignments based on sound classes**. In Slavkovik, M. and D. Lassiter (eds.): *New directions in logic, language, and computation*. Springer: Berlin and Heidelberg. 2012b; 32–51.
Publisher Full Text

List JM: **Sequence Comparison in Historical Linguistics**. Düsseldorf: Düsseldorf University Press. 2014.
Publisher Full Text

List JM: **Network Perspectives on Chinese Dialect History.** *Bulletin of Chinese Linguistics.* 2015; **8**: 42–67.
Reference Source

List JM: **Beyond Cognacy: Historical Relations Between Words and Their Implication for Phylogenetic Reconstruction.** *J Lang Evol.* 2016; **1**(2): 119–36.
Publisher Full Text

List JM: **A Web-Based Interactive Tool for Creating, Inspecting, Editing,**

and Publishing Etymological Datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Association for Computational Linguistics. 2017; 9–12.
Publisher Full Text

List JM: **Automated methods for the investigation of language contact with a focus on lexical borrowing.** *Lang Linguist Compass.* 2019a; **13**(10): e12355.
Publisher Full Text

List JM: **Automatic Inference of Sound Correspondence Patterns Across Multiple Languages.** *Comput Linguist.* 2019b; **45**(1): 137–61.
Publisher Full Text

List JM: **Die Bedeutung der Grundline.** [The importance of the baseline]. Von Wörtern und Bäumen. 2019c; **3**(1): 902.
Reference Source

List JM: **EDICTOR. a Web-Based Tool for Creating, Editing, and Publishing Etymological Datasets. Version 2.0.0**. (version 2.0.0). Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Reference Source

List JM, Anderson C, Tresoldi T, *et al.*: **Cross-Linguistic Transcription Systems. Version 2.1.0**. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021b.
Publisher Full Text

List JM, Forkel R: **LingRex: Linguistic Reconstruction with LingPy.** (version 1.1.1). Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021a.
Publisher Full Text

List JM, Forkel R: **CLDF dataset accompanying List and Forkel's "Borrowing Detection in Multilingual Wordlists" from 2021.** (Version v1.1) [Data set]. *Zenodo.* 2021b.
http://www.doi.org/10.5281/zenodo.5037101

List JM, Greenhill SJ, Anderson C, *et al.*: **CLICS^2: An Improved Database of Cross-Linguistic Colexifications Assembling Lexical Data with Help of Cross-Linguistic Data Formats.** *Linguistic Typology.* 2018; **22**(2): 277–306.
Publisher Full Text

List JM, Greenhill S, Gray RD: **The potential of automatic word comparison for historical linguistics.** *PLoS One.* 2017; **12**(1): e0170046.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

List JM, Lopez P, Bapteste E: **Using Sequence Similarity Networks to Identify Partial Cognates in Multilingual Wordlists**. In *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*. Berlin: Association of Computational Linguistics. 2016; 599–605.
Publisher Full Text

List JM, Nelson-Sathi S, Geisler H, *et al.*: **Networks of Lexical Borrowing and Lateral Gene Transfer in Language and Genome Evolution.** *Bioessays.* 2014a; **36**(2): 141–50.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

List JM, Shijulal NS, Martin W, *et al.*: **Using Phylogenetic Networks to Model Chinese Dialect History.** *Language Dynamics and Change.* 2014b; **4**(2): 222–52.
Publisher Full Text

List JM, Rzymski C, Greenhill S, *et al.*: **Concepticon. A Resource for the Linking of Concept Lists. Version 2.5.0 (version 2.5.0)**. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021a.
Publisher Full Text

Liú L, Wáng H, Bǎi Y: **Xiàndài Hànyǔ Fāngyán Héxīncí, Tèzhēng Cíjí**. Nánjīng 南京: Fènghuáng 凤凰. 2007.

Mennecier P, Nerbonne J, Heyer E, *et al.*: **A Central Asian Language Survey.** *Language Dynamics and Change.* 2016; **6**(1): 57–98.
Publisher Full Text

Montenegro Á, Avis C, Weaver A: **Modeling the Prehistoric Arrival of the Sweet Potato in Polynesia.** *J Archaeol Sci.* 2008; **35**: 355–67.
Publisher Full Text

Nakhleh L, Ringe D, Warnow T: **Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages.** *Language.* 2005; **81**(2): 382–420.
Publisher Full Text

Nelson-Sathi S, List JM, Geisler H, *et al.*: **Networks Uncover Hidden Lexical Borrowing in Indo-European Language Evolution.** *Proc Biol Sci.* 2011; **278**(1713): 1794–1803.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Onysko A: **Reconceptualizing Language Contact Phenomena as Cognitive Processes.** Ed. by E. Zenner, A. Backus, and E. Winter-Froemel, 2019; 23–50.
Publisher Full Text

Pritchard JK, Stephens M, Donnelly P: **Inference of Population Structure Using Multilocus Genotype Data.** *Genetics.* 2000; **155**(2): 945–59.
PubMed Abstract | **Free Full Text**

Prokić J, Wieling M, Nerbonne J, *et al.*: **Multiple sequence alignments in linguistics.** In *Proceedings of the EACL 2009 Workshop on Language Technology Social Sciences, Humanities, and Education*. 2009; 18–25.
Publisher Full Text

Sagart L, Jacques G, Lai Y, *et al.*: **Dated Language Phylogenies Shed Light on the Ancestry of Sino-Tibetan.** *Proc Natl Acad Sci U S A.* 2019; **116**(21): 10317–22.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Schleicher A: **Die Darwinsche Theorie Und Die Sprachwissenschaft: Offenes Sendschreiben an Herrn Dr. Ernst Haeckel.** Weimar: Hermann Böhlau, 1863.
Reference Source

Schweikhard NE, List JM: **Developing an annotation framework for word formation processes in comparative linguistics.** *SKASE Journal of Theoretical Linguistics.* 2020; **17**(1); 2–26.
Publisher Full Text

Swadesh M: **Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos.** *Proceedings of the American Philosophical Society.* 1952; **96**(4): 452–63.
Reference Source

Swadesh M: **Towards Greater Accuracy in Lexicostatistic Dating.** *International Journal of American Linguistics.* 1955; **21**(2): 121–37.
Publisher Full Text

Szeto PY, Yurayong C: **Establishing a Sprachbund in the Western Lingnan region: Conceptual and methodological issues.** *Folia Linguistica.* forthcoming.

Tadmor U: **Loanwords in the World's Languages: Findings and Results.** In *Loanwords in the World's Languages: A Comparative Handbook.* edited by Martin Haspelmath and Uri Tadmor, Berlin; New York: de Gruyter, 2009; 55–75.
Publisher Full Text

van der Ark R, Mennecier P, Nerbonne J, *et al.*: **Preliminary Identification of Language Groups and Loan Words in Central Asia**. In *Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons*. 2007; 13–20.
Reference Source

Wang F: **Language Contact and Language Comparison. The Case of Bai.** PhD, Hong Kong: City University of Hong Kong, 2004.
Reference Source

Wu MS, List JM: **Annotating Cognates in Phylogenetic Studies of South-East Asian Languages.** *Humanities Commons.* Preprint, currently under review. 2021.
Publisher Full Text

Wu MS, Schweikhard NE, Bodt TA, *et al.*: **Computer-Assisted Language Comparison: State of the Art.** *Journal of Open Humanities Data.* 2020; **6**(1): 2.
Publisher Full Text

Zhang L, Manni F, Fabri R, *et al.*: **Detecting Loan Words Computationally.** In: *Variation Rolls the Dice. A Worldwide Collage in Honour of Salikoko S. Mufwene*. Amsterdam: Benjamins. forthcoming. 269–288.
Reference Source

Open Peer Review

Current Peer Review Status: ✓ ✓ ✓ ✓

Version 2

Reviewer Report 02 September 2021

https://doi.org/10.21956/openreseurope.15171.r27472

✓ **John Nerbonne**
[1] Computational Linguistics and chair of Humanities Computing, University of Groningen, Groningen, The Netherlands
[2] Computational Linguistics and chair of Humanities Computing, University of Groningen, Groningen, The Netherlands

I find this much clearer, and I approve of its indexing without reservation.

It's a strong paper about an important topic.

I noted a couple of typos, problems in font embedding, and slips in diction, however.

p.5 "requires to search" => requires one to search"
p.8 "borrowing sare" => borrowings are
p.11 "considerably small" => rather small
p.18, phon. transcription not rendered in .pdf

Finally, I would have appreciated some speculation on WHERE the improvement comes from, but I wouldn't suggest holding anything up for that reason.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 02 Sep 2021

Johann-Mattis List

We are very grateful for the assessment by the reviewer and will try to submit a third version of our article soon which makes up for all points raised.

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 12 August 2021

https://doi.org/10.21956/openreseurope.14926.r27264

✓ **Kenichi W. Okamoto**

[1] Department of Biology, University of St. Thomas, St. Paul, MN, USA
[2] Department of Biology, University of St. Thomas, St. Paul, MN, USA
[3] Department of Biology, University of St. Thomas, St. Paul, MN, USA

This manuscript introduces and illustrates a new technique for identifying borrowing events found both within and across multiple language families. While computational approaches are a well-suited and increasingly popular tool for exploring language change when the underlying lexical data are likely orthologous, linguistic borrowing presents a major and very real complication. For many languages the phenomenon is often poorly attested in the historical record. It is sufficiently common to potentially bias several standard tools from systematic biology for assessing phylogenies. These challenges are exacerbated by the difficulty of manually assessing the evolutionary fate of proto-words across several languages to distinguish xenologs from orthologs; doing so requires expertise in the idiosyncratic evolution of several, often unrelated lineages.

Here, the authors aim to make some headway viz. the difficulty of identifying borrowing events when multiple language families are involved. Briefly, the authors propose a two-step procedure, first identifying cognates within language families and then seeking to identify cognates across families. The authors then provide a case study and examine patterns of lexical borrowing within a well-studied Sprachbund of three language families in southeast China. The soundness of the proposed approach is assessed through well-established text clustering methods. They then introduce several techniques by which comparative linguists might begin to quantitatively characterize how language contact develops.

General comments:

I find this last aspect of the study particularly useful, as it moves the study of Areal linguistics in a productive direction. For instance, their use of admixture plots not only allows linguists to visualize patterns of borrowings spatially; it isn't hard to envision extensions that can map similar measures onto features of physical geography (for instance) in a way that can yield insights previously hard to explore statistically. These approaches have proven a boon to our understanding of how gene flow interacts with ecological processes in biogeography. Whether such extensions in linguistics will materialize remains to be seen, but on balance my sense is the applications described here lay a plausible point of departure for quantitatively and I think usefully comparing language contact dynamics across Sprachbunds.

Specific comments:

Background:
It wasn't clear what was meant by "shortcut status" here.

Materials:
Second paragraph, "Having shown..." the authors note "the dialect data do not provide strict information on actual words" - it was admittedly unclear to me why this was so. How plausible is it that the phonetic properties of the morphemes/characters change sufficiently to drown out the cross-dialectical signal in the context of different words? Admittedly, I'm not familiar enough with subtle morpheme distinctions in Sino-Tibetan to have a good feel about this, but maybe some clarification/references on this point would help.

Methods:
Second paragraph: "For our specific use case..." versus "Since one would expect, ...". I wondered if this contrast means that the use of full cognates should be the more parsimonious starting point when applying the approach to other contexts, or if the search for partial cognates makes more sense? Presumably there is some degree of compounding in the morphology of many languages that complicate (esp. automated) attempts at identifying borrowing patterns. Could the authors comment on the advisability of using partial rather than full cognates during the first step beyond their case study?

Third paragraph: "If the average distance ..." are there any rough rules of thumb on reasonable thresholds on which the authors can comment? Particularly perhaps based on their manual comparisons later.

"Specific Results":
Although I didn't have immediate access to List 2014, presumably the "True chain" (sensu Bagga and Baldwin 1998) consists of the manually annotated cognates/borrowings (analogously with the F-statistic's True/False positives)? I suspect a quick clarification would help. Also, unlike the rest of the manuscript, this segment alone appears written in the first person.

Table 2: Since it won't be at all computationally costly, it might be interesting to see what a bootstrapped or other resampling analysis of the evaluation metrics would show viz. characterizing the variability around the test statistics.

Third paragraph: "In order to test...": I may have missed something in the description of the resampling test here, but a caveat is that this analysis involves multiple comparisons across two lists (Swadesh/Leipzig-Jakarta), so the significance test threshold mentioned at the end likely can be adjusted. Granted, the proportion is still quite low for both lists even when doing so, which I suspect is still linguistically meaningful.

Finally, as I am not particularly familiar with the three language families analyzed, I cannot comment on how reasonable the present study's findings will strike a specialist in the language families explored. I have, however, studied the underlying workflow code, and the authors' implementations appear technically sound as far as I can tell.

References

1. Bagga A, Baldwin, B: Entity-based cross-document coreferencing using the Vector Space Model. *ACL '98/COLING '98*. 1998; **1**: 79-85 Publisher Full Text

Is the work original in terms of material and argument?

Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?

Yes

Is the work clearly and cogently presented?

Yes

Is the argument persuasive and supported by evidence?

Yes

If any, are all the source data and materials underlying the results available?

Yes

Does the research article contribute to the cultural, historical, social understanding of the field?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary biology, computational biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 12 Aug 2021

Johann-Mattis List

We thank the reviewer for taking the time for this very thorough review. What is specifically nice is that the reviewer -- due to the different background -- expands the perspectives on our study even more, providing some new suggestions that will help us to improve the study further.

Competing Interests: No competing interests were disclosed.

Author Response 19 Aug 2021

Johann-Mattis List

Dear Kenichi W. Okamoto, The revised version of the manuscript which we just submitted hopefully addresses the points you raised in your review. We list out individual responses to the points in bullet point form below.

- *It wasn't clear what was meant by "shortcut status" here.*
  We have clarified this now.

- *Table 2: Since it won't be at all computationally costly, it might be interesting to see what a bootstrapped or other resampling analysis of the evaluation metrics would show viz. characterizing the variability around the test statistics.*
  We have expanded the evaluation measures in the general analysis now drastically, and this has helped us to clarify for ourselves what the particular benefits of the method are. Especially the introduction of the baselines proved important. For bootstrapping or similar evaluations, we were not sure how to proceed in concrete, but hope that our new tests are satisfying enough.

- *Although I didn't have immediate access to List 2014, presumably the "True chain" (sensu Bagga and Baldwin 1998) consists of the manually annotated cognates/borrowings (analogously with the F-statistic's True/False positives)? I suspect a quick clarification would help. Also, unlike the rest of the manuscript, this segment alone appears written in the first person.*
  We have extended this part, not in detail, but by showing what the measures do. We refer to a page range in the book by List (2014). Since the book is open access, readers interested in understanding these evaluation measures can now find all the necessary information needed.

- *Third paragraph: "In order to test...": I may have missed something in the description of the resampling test here, but a caveat is that this analysis involves multiple comparisons across two lists (Swadesh/Leipzig-Jakarta), so the significance test threshold mentioned at the end likely can be adjusted. Granted, the proportion is still quite low for both lists even when doing so, which I suspect is still linguistically meaningful.*
  We discussed if resampling would apply, but have to admit that it is not entirely clear to us, if this applies in our examples, where we merely compare significance scores among two different datasets and do not seek to find overall significant results. We try to avoid to fix a significance value (which would, if we understand Bonferroni correction well, be 0.025 due to the fact that we have two tests), so we just left out this part, where we mention the 5%, which are anyway disputed among scholars. We hope that the major message here (that Swadesh (1955) splits the dataset with the automated borrowings more meaningful than the Leipzig-Jakarta list) is still correct in this form, even if we avoid to open a discussion about significance values.

Competing Interests: No competing interests were disclosed.

Reviewer Report 06 August 2021

https://doi.org/10.21956/openreseurope.14926.r27266

✓

George S. Starostin

[1] Institute of Oriental and Classical Studies, Russian State University for the Humanities, Moscow, Russian Federation

[2] Institute of Oriental and Classical Studies, Russian State University for the Humanities, Moscow, Russian Federation

[3] Institute of Oriental and Classical Studies, Russian State University for the Humanities, Moscow, Russian Federation

The paper is a serious step forward in the development of new techniques to automatically identify potential borrowings between groups of (assumedly) unrelated languages. The new method follows the authors' earlier research on similar algorithms for eliciting actual cognates between (assumedly) related languages and is largely operating along the same lines. The selected data set upon which the algorithm has been tested is valid, though in the future it would be essential to expand the test base by data from other areas of the world to confirm its universal applicability.

I have detected one serious (though local) mistake in the main bulk of the paper: the MC reconstruction for the word 'face' is given as *mjuwk, when in reality that is the reconstruction for 'eye' (目 mu); the proper MC reconstruction for 'face' is *mjienH, which is, of course, the form that is phonetically compatible with both the modern Sinitic reflexes and the borrowed HM forms. This should by all means be corrected, and perhaps the other MC forms in the data should be double-checked as well.

All in all, the work definitely merits indexing.

Is the work original in terms of material and argument?

Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?

Yes

Is the work clearly and cogently presented?

Yes

Is the argument persuasive and supported by evidence?

Yes

If any, are all the source data and materials underlying the results available?
Yes

Does the research article contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Comparative-historical linguistics, Chinese historical linguistics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 06 Aug 2021

Johann-Mattis List

We thank the reviewer a lot for the review, and specifically for catching the indeed serious if not even embarrassing error regarding the pronunciation of Middle Chinese "face", which we will of course correct. This shows how important it is when dealing with multi-disciplinary topics that reviewer experts from different fields check a study for particular problems.

Competing Interests: No competing interests were disclosed.

Author Response 19 Aug 2021

Johann-Mattis List

Dear George S. Starostin, Our revised version now lists the correct form for "face". We are very thankful that you caught this mistake. When looking up the character reading in Baxter (1992), we must have strangely swapped the characters, but this has been corrected now.

Competing Interests: No competing interests were disclosed.

Reviewer Report 30 July 2021

https://doi.org/10.21956/openreseurope.14926.r27263

?

John Nerbonne
[1] Computational Linguistics and chair of Humanities Computing, University of Groningen,

Groningen, The Netherlands
[2] Computational Linguistics and chair of Humanities Computing, University of Groningen, Groningen, The Netherlands
[3] Computational Linguistics and chair of Humanities Computing, University of Groningen, Groningen, The Netherlands

Summary
The paper proposes a two-step process for identifying borrowings. First, one applies earlier work by the authors to obtain language-family-internal cognate sets, and second, it applies a sound-class alignment algorithm (SCA) due to the first author. The paper evaluates using a substantial data set and reports F1-scores between 0.86 (for detecting borrowing) and 0.88 (for detecting cognacy).

Comments
The paper reports on work advancing the detection of loan words, which is an important issue, and it seems to be successful, so I recommend that it be indexed, but I urge the authors to consider the remarks below. The use of the admixture plots in Fig. 3 is a clear advance, and Table 3, noting the relatively low proportion of non-borrowed words in the Swadesh list, will be extremely interesting to all historical linguists!

I had the advantage of reading Pui Yiu Szeto's review, which is fortunate since I would not presume to comment on the quality of the work with respect to Sinology. But where Szeto praises the article's accessibility, I would urge the authors to make the work even more easily accessible to specialists and non-specialists.

Most of my complaints concern clarity, and given the (laudable!) availability of both data and software, might not bother readers who consult all the available material, but there's something to be said for providing information in a relatively self-contained paper.

I assume that input data is the usual table of concepts (rows) and varieties/doculects (columns) (or a 90° rotation of that), but it would help to be explicit. It is unclear to me whether input language data is annotated with respect to its putative language family (going beyond the usual input data), as Fig. 2 suggests. Identifying families is often part of the task, of course.

The notion 'partial cognate' is crucial, but neither defined nor illustrated. I'd suggest doing both, even if this repeats material from List et al. 2016). In the same vein, it would help to illustrate the two steps on a small data set. These two points should be addressed.

The clarity of the paper might also improve if the authors would explain (and illustrate) step one thoroughly and then turn to step two (in the "Methods" section). As it now stands, the first para. explains both steps, while the second returns to the first step. Page 4 reports that partial cognates were converted to full cognates (in the first step) using a variety of the LexStat algorithm as modified recently in (Wu & List submitted) that "assigns partial cognates to full cognate sets which have at least one cognate set in common". The idea is to recover cognate morphemes, even when these only occur within words, e.g. Dutch *moestuin* /'mu:s.tœyn/ 'vegetable garden' is compared with German *Gemüse* /gə.'my:.zə/ 'vegetable', which share the morpheme {mUS} (an example I could think of quickly, but if the authors wish to illustrate the concept, as I suggest, then using

Chinese data would probably be preferable. While I have a vague idea what the technique must be doing (a sort of Longest Common Subsequence algorithm?), it would help to at least illustrate it concretely, especially since the cited source still awaits publication, and since its focus seem to be on annotation, not (partial) cognate detection.[1] I wouldn't mind reading a dataflow diagram with example data.

The second step uses List's SCA algorithm, as noted in the summary, which tests pronunciation similarity with respect to the question of common genealogy, asking whether two words are likely to have descended from the same source – either via ancestry or borrowing. All word pairs within putative cognate sets are tested and the sets are identified as involving borrowing the average distance falls below a user-defined threshold. This is novel, but I would urge the authors to speculate on whether this is the source of the improvement they demonstrate. Comparing all pairs of words based on SCA distance would follow an obvious step from earlier work (Zhang *et al* 2021), but the advantage of the approach in this paper might be in comparing cognate *sets* with respect to similarity rather than just individual words from the two different languages. It would be good to hear from the authors, especially if they are of a different opinion. see below.

It is also unclear how the "user-defined threshold" (p.4) is determined. These issues should be clarified, since they are relevant with respect to how independently the procedures work (independent of annotation, and independent of user-supplied parameters), and how generally applicable they can be.

The authors are candid about their techniques not indicating the direction of borrowing, but I assume that it will be difficult to detect cases where two languages borrow from a third, such as the word 'sputnik' in many languages. The issue might be made explicit.

The sensitivity to phonotactics (noted on p.10) is interesting, since the example is the sort that edit distance approaches have little difficulty with. Perhaps the authors would care to comment further.

The authors might wish to refer to neighbor net and splits-tree as potentially useful ways of analyzing the historical data once borrowings have been determined.

Form: The paper is very readable, but there are a couple of slips.

- withour => without (caption, Tbl.1)
- concepts was selected => were (p.4, l.1)
- amount of words, etc. => number of words (passim); more conservative readers use 'amount' only with mass nouns. So this may be a reactionary recommendation.
- Zhang et al. now has page numbers: 269-288 and has been announced for appearance in Oct. (References)

Conclusion: The paper should be indexed after the requests for clarifying remarks have been addressed. These concern (i) what partial cognates are; (ii) how the revised LexStat algorithm works, including an example application on illustrative data; and (iii) whether comparing average distances between potential cognate sets is the source of the improvement in detection.

[1] I looked at Wu & List (submitted) without reading it closely, but it didn't clarify this point for me. Perhaps the technique might be identified more specifically via page numbers or name.

Is the work original in terms of material and argument?
Partly

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
Yes

Is the work clearly and cogently presented?
Partly

Is the argument persuasive and supported by evidence?
Yes

If any, are all the source data and materials underlying the results available?
Yes

Does the research article contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational linguistics, comparative linguistics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 04 Aug 2021

Johann-Mattis List

We thank the reviewer a lot for taking the time for this very thoughtful and very detailed report. We will try to address these concerns soon after we have received one more review report which may point to additional problems in our approach which have not been covered so far.

Competing Interests: No competing interests were disclosed.

Author Response 19 Aug 2021

Johann-Mattis List

Dear John Nerbonne, we have now revised the manuscript and submitted an updated version, in which we tried to account for all changes you suggested. Below, we list our modifications in the form of bullet points.

- ○ *I assume that input data is the usual table of concepts (rows) and varieties/doculects*

(columns) (or a 90° rotation of that), but it would help to be explicit. It is unclear to me whether input language data is annotated with respect to its putative language family (going beyond the usual input data), as Fig. 2 suggests. Identifying families is often part of the task, of course.
We have added a workflow figure in which we hope that the algorithms are clarified sufficiently now.

- *The notion 'partial cognate' is crucial, but neither defined nor illustrated. I'd suggest doing both, even if this repeats material from List et al. 2016). In the same vein, it would help to illustrate the two steps on a small data set. These two points should be addressed.*
  We hope our Figure 2 will address this question now, we also add clarification in Figure 3, and we refer to additional literature. We emphasize that while we could provide many more examples in this study, we consider that it would lead too far away if we had to comment on all ongoing debates on partial cognates. For this reason, we now provide more examples but also refer to the literature which is in our impression crucial.

- *This is novel, but I would urge the authors to speculate on whether this is the source of the improvement they demonstrate. Comparing all pairs of words based on SCA distance would follow an obvious step from earlier work (Zhang et al 2021), but the advantage of the approach in this paper might be in comparing cognate *sets* with respect to similarity rather than just individual words from the two different languages. It would be good to hear from the authors, especially if they are of a different opinion. see below.*
  We explicitly comment on this now and contrast our approach with an earlier approach where no direct evaluation was, however, provided.

- *It is also unclear how the "user-defined threshold" (p.4) is determined. These issues should be clarified, since they are relevant with respect to how independently the procedures work (independent of annotation, and independent of user-supplied parameters), and how generally applicable they can be.*
  We discuss the difficulties of finding thresholds in our methods section now.

- *The authors are candid about their techniques not indicating the direction of borrowing, but I assume that it will be difficult to detect cases where two languages borrow from a third, such as the word 'sputnik' in many languages. The issue might be made explicit.*
  The method can detect all those cases where cognates spread several language families. If sputnik only occurred in Hmong-Mien languages, but not in Tai-Kadai and Chinese, we would not be able to detect the word, but if it occurs across two or more language families, we should be able to find it.

- *The sensitivity to phonotactics (noted on p.10) is interesting, since the example is the sort that edit distance approaches have little difficulty with. Perhaps the authors would care to comment further.*
  We address this now explicitly in the results section, emphasizing that it is the matching of vowels and consonants, which is problematic here. We call this "phonotactics", which may be missleading, but we hope that our additional statement clarifies the problem.

- *withour => without (caption, Tbl.1)*
  We corrected this, thanks.

- *concepts was selected => were (p.4, l.1)*
  We corrected this, thanks.

- *amount of words, etc. => number of words (passim); more conservative readers use ‘amount’ only with mass nouns. So this may be a reactionary recommendation*.
  We corrected this, thanks.

- *Zhang et al. now has page numbers: 269-288 and has been announced for appearance in Oct. (References)*
  We corrected this, thanks.

- *These concern (i) what partial cognates are;*
  We hope, our response above and the revised manuscript make this clearer now.

- *(ii) how the revised LexStat algorithm works, including an example application on illustrative data; and*
  We also hope to have clarified this point.

- *(iii) whether comparing average distances between potential cognate sets is the source of the improvement in detection.*
  We hope our additional statistics on thresholds provide the necessary insights here.

Competing Interests: No competing interests were disclosed.

Reviewer Report 21 July 2021

https://doi.org/10.21956/openreseurope.14926.r27265

✓ **Pui Yiu Szeto**

[1] Department of Linguistics, The University of Hong Kong, Hong Kong, China
[2] Department of Linguistics, The University of Hong Kong, Hong Kong, China
[3] Department of Linguistics, The University of Hong Kong, Hong Kong, China

This article introduces a new automated method for detecting lexical borrowing in contact scenarios involving multiple languages of different genetic affiliations. Although there have been a number of studies on automated detection of lexical borrowing over the past decade, this study represents a major breakthrough since this is the first attempt to apply this approach to more than two languages at the same time.

Overall speaking, this is a very informative and well-organized article. The authors manage to provide a clear and concise introduction to the background information of the subject matter. The methodology part of studies concerning computational methods often looks rather daunting to non-specialists, but this article presents everything in a reader-friendly manner, while providing the relatively technical details in the supplementary materials for those who are interested in replicating the experiments or testing the method on a different set of data. Granted, like other automated methods in computational historical linguistics, the results generated by this new method must be checked by historical linguists with specialist knowledge of the languages concerned; nonetheless it appears to be a very handy tool for the study of lexical borrowing in contact scenarios.

There are some minor issues which the authors may want to take into consideration when they carry out further studies along this direction.

1. Reading the abstract, one will normally expect that languages spoken in Southeast Asian countries like Vietnam, Cambodia, Thailand, and Laos are considered in this study. However, all the languages involved are actually spoken in Southern China. Apart from being potentially misleading, a problem of exclusively focusing on languages in Southern China is that there may not be sufficient language-family-internal variation in terms of borrowing patterns to yield interesting results, as shown in the admixture plots of Tai-Kadai and Hmong-Mien languages in Figure 3. Given the considerable structural differences between Tai-Kadai languages spoken inside and outside China due to varying degrees of Sinitic influence (Szeto and Yurayong [forthcoming]), it will be interesting to incorporate at least a couple of Tai-Kadai languages spoken outside China to see whether they display a remarkably different pattern of lexical borrowing. If this is the case, it may suggest that the method can even help access the intensity of language contact.

2. The authors rightly point out that the Chinese dialect data in this study is less than ideal. First, there are many missing concepts in some of the dialects. More importantly, the data comes from various sources, subject to non-uniform (and sometimes questionable) elicitation processes. A case in point is the word form for 'correct'. In Yue dialects , the most common form is *ŋam* 啱, which is possibly a loanword of Tai-Kadai origin (Li 1990). I suggest using the *Great Dictionary of Modern Chinese Dialects* (Li 2002) as the sole source of Chinese dialect data. Li (2002) is a compendium of dictionaries for 42 Chinese dialects following a uniform format. This can surely improve the accuracy and coverage rate of the Chinese dialect data.

References

1. Szeto, Pui Yiu Chingduang, Yurayong: Establishing a Sprachbund in the Western Lingnan region: Conceptual and methodological issues source. (forthcoming).
2. [Li, Jinfang] 李锦芳: 粤语中的壮侗语族语言底层初析 [A preliminary analysis of Tai-Kadai substrate words in Yue]. 中央民族学院学报 [*Journal of the Central Institute for Nationalities]*. 1990. 71-76.
3. [Li, Rong] 李榮 (ed.): 現代漢語方言大詞典 [The great dictionary of Modern Chinese dialects]. *Jiangsu Education Publishing House*. 2002.

Is the work original in terms of material and argument?
Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
Yes

Is the work clearly and cogently presented?
Yes

Is the argument persuasive and supported by evidence?
Yes

If any, are all the source data and materials underlying the results available?
Yes

Does the research article contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Language contact, linguistic typology, language change, East and Southeast Asian languages

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 27 Jul 2021

Johann-Mattis List

We thank the reviewer very much for these very valuable and also inspiring comments. We will respond to them in more detail later, when more reviews have been received and also try to account for the suggestions in an updated version of our manuscript.

Competing Interests: No competing interests were disclosed.

Author Response 19 Aug 2021

Johann-Mattis List

Dear Pui Yiu Szeto, we have now had time to revise the manuscript and submitted a new version. Below, we list your suggestions to improve the study and add our answers.

- *Reading the abstract, one will normally expect that languages spoken in Southeast Asian countries like Vietnam, Cambodia, Thailand, and Laos are considered in this study. However, all the languages involved are actually spoken in Southern China.*

We have made sure to be more precise in the new version, specifying Southern China.

- *Given the considerable structural differences between Tai-Kadai languages spoken inside and outside China due to varying degrees of Sinitic influence (Szeto and Yurayong [forthcoming]), it will be interesting to incorporate at least a couple of Tai-Kadai languages spoken outside China to see whether they display a remarkably different pattern of lexical borrowing.*
  We now mention the idea to compare languages inside and outside of China as a potential case study in our conclusion.

- *I suggest using the Great Dictionary of Modern Chinese Dialects (Li 2002) as the sole source of Chinese dialect data. Li (2002) is a compendium of dictionaries for 42 Chinese dialects following a uniform format. This can surely improve the accuracy and coverage rate of the Chinese dialect data.*
  We added a paragraph in the methods section, in which we critically assess the potential of including the dictionary series by Lǐ (2002) and conclude that due to the way the data was gathered, it would unfortunately be very difficult (if not impossible) to extract the information in the formats needed for our work.

Competing Interests: No competing interests were disclosed.