Using Sequence Similarity Networks to Identify Partial Cognates in Multilingual Wordlists

Johann-Mattis List
CRLAO/UPMC
2 rue de Lille
75007 Paris
mattis.list@lingpy.org

Philippe Lopez
UPMC
9 quai de Bernard
75005 Paris
philippe.lopez@upmc.fr

Eric Bapteste
UPMC
9 quai de Bernard
75005 Paris
eric.bapteste@upmc.fr

Abstract

Increasing amounts of digital data in historical linguistics necessitate the development of automatic methods for the detection of cognate words across languages. Recently developed methods work well on language families with moderate time depths, but they are not capable of identifying cognate morphemes in words which are only partially related. Partial cognacy, however, is a frequently recurring phenomenon, especially in language families with productive derivational morphology. This paper presents a pilot approach for partial cognate detection in which networks are used to represent similarities between word parts and cognate morphemes are identified with help of state-of-the-art algorithms for network partitioning. The approach is tested on a newly created benchmark dataset with data from three sub-branches of Sino-Tibetan and yields very promising results, outperforming all algorithms which are not sensible to partial cognacy.

1 Introduction

In a very general notion, cognacy is similar to the concept of *homology* in biology (Haggerty et al. 2014), denoting a relation between words which share a common history (List 2014b). In classical linguistics, borrowings are often excluded from this notion (Trask 2000). Quantitative approaches additionally distinguish cognates which have retained, and cognates which have shifted their meaning (Starostin 2013b). Further aspects of cognacy are rarely distinguished, although they are obvious and common. Words which go back to the same ancestor form can for example have been morphologically modified, such as French *soleil* which does not go directly back to Latin *sōl* `sun' but to *sōliculus* `small sun' which is itself a derivation of *sōl* (Meyer-Lübke 1911).

Variety	Form	Character	Cognacy
Fúzhōu	ŋuoʔ^{5}	月	1
Měixiàn	ŋiat^{5} kuoŋ^{44}	月光	1 2
Wēnzhōu	ɳy^{21} kuɔ^{35} vai^{13}	月光佛	1 2 3
Běijīng	yɛ^{51} liɑŋ^{1}	月亮	1 4

Table 1: Partial cognacy in Chinese dialects.

Another problem are words which have been created from two or more morphemes via processes of *compounding*. While these cases are rather rare in the core vocabulary of Indo-European languages, they are very frequent in South-East Asian language families like Sino-Tibetan or Austro-Asiatic. In 200 basic words across 23 Chinese dialects (Ben Hamed and Wang 2006), for example, almost 50% of the nouns and more than 30% of all words consist of two or more morphemes (see the Sup. Material for details).

The presence of words consisting of more than one morpheme challenges the notion that words can either be cognate or not. It poses problems for phylogenetic approaches which require binary presence-absence matrices as input and model language evolution as cognate gain and cognate loss (Atkinson and Gray 2006). This is illustrated in Table 1 where words for `moon' in four Chinese dialects (Hóu 2004) are compared, with cognate elements being given the same color. If we assign cognacy *strictly*, only matching those words which are identical in all their elements (Ben Hamed and Wang 2006), we would have to label all words as being not cognate. If we assign cognacy *loosely* (Satterthwaite-Phillips 2011), labeling all words as cognate when only they share a common morpheme, we would have to label all

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 599–605,
Berlin, Germany, August 7-12, 2016.

words as cognate. No matter how we code in phylogenetic analyses, as long as we use binary states, we will loose information (List 2016).

Partial cognacy is also a problem for current cognate detection algorithms which compare words in their entirety (List 2014b, Turchin et al. 2010). Given the frequency of compound words in South-East Asian languages, it is not surprising that the algorithms perform much worse on diverse South-East Asian language families, than they perform on other language families where compounding is less frequent (List 2014b:197f).

This paper presents a new algorithm for cognate detection which does not identify cognate *words* but instead searches for cognate *elements* in words. The algorithm takes multilingual word lists as input and outputs statements regarding the cognacy of morphemes, just as the ones shown in the last column of Table 1, where identical numerical IDs are given for all morphemes identified as cognate.

Dataset	**Bai**	**Chinese**	**Tujia**
Languages	9	18	5
Words	1028	3653	513
Concepts	110	180	109
Strict Cogn.	285	1231	247
Partial Cogn.	309	1408	348
Sounds	94	122	57
Source	Wang, 2006	Běijīng Dàxué, 1964	Starostin, 2013b

Table 2: Partial cognate detection gold standard

2 Materials

Three gold standard datasets from different branches of Sino-Tibetan with different degrees of diversity were prepared, including Bai dialects, Chinese dialects, and Tujia dialects. All datasets were taken from existing datasets with cognate codings provided independently. To facilitate further use of the data, all languages were linked to Glottolog (Hammarström et al. 2015) and all concepts were linked to the Concepticon (List et al. 2016a). Furthermore, phonetic transcriptions were cleaned by segmenting phonetic entries into meaningful sound units and unifying phonetic variants representing the same pronunciation. Morphological segmentation was not required, since all languages in our sample (and the majority of all South-East Asian languages) have a morpheme-syllabic structure in which each syllable denotes one morpheme. Partial cognate judgments are displayed with help of multiple integer IDs assigned to a word in the order of its morphemes, as displayed above in Table 1. For the Chinese dataset, partial cognate information was provided in the source itself, for Bai and Tujia, it was manually derived from the cognate judgments in the sources. Detailed information regarding the datasets is given in Table 2, and the full dataset along with further information is given in the Sup. Material.

3 Methods

The workflow for partial cognate detection consists of three major steps. (1) In a first step, pairwise sequence similarities are determined between all morphemes of all words in the same meaning slot in a word list. (2) These similarities are then used to create a similarity network in which nodes represent morphemes and edges between the nodes represent similarities between the morphemes. (3) In a third step, an algorithm for network partitioning is used to cluster the nodes of the network into groups of cognate morphemes.

3.1 Sequence Similarity

There are various ways to determine the similarity or distance between words and morphemes. A general distinction can be made between *language-independent* and *language-specific* approaches. The former determine the word similarity independently of the languages to which the words belong. As a result, the scores only depend on the substantial and structural differences between words. Examples for language-independent similarity measures are SCA distances, as produced by the Sound-Class-Based Phonetic Alignment algorithm (List 2012b), or PMI similarities as produced by the Weighted String Alignment algorithm (Jäger 2013). Language-specific approaches, on the other hand, are based on previously identified recurring correspondences between the languages from which the words are taken (List 2014b: 48-50) and may differ across languages.[1] An example for language-specific similarity measures is the LexStat algorithm, first proposed in List (2012a) and later refined in List

[1]Comparing, for example, German *Kuckuck* with French *coucou* and English *cuckoo* may yield quite different scores, although the English and the French words are almost identical in pronunciation.

		A						
Fúzhōu	ŋuoʔ^{5}							
Měixiàn	ŋiat^{5}	0.44						
	kuoŋ^{44}	0.78	0.78					
Wēnzhōu	ɳʚy^{21}	0.30	0.35	0.67				
	kuɔ^{35}	0.80	0.85	0.27	0.67			
	vai^{13}	0.85	0.85	0.82	0.73	0.73		
Běijīng	yɛ^{51}	0.77	0.84	0.73	0.56	0.56	0.66	
	liɑŋ^{1}	0.78	0.78	0.44	0.67	0.82	0.82	0.80

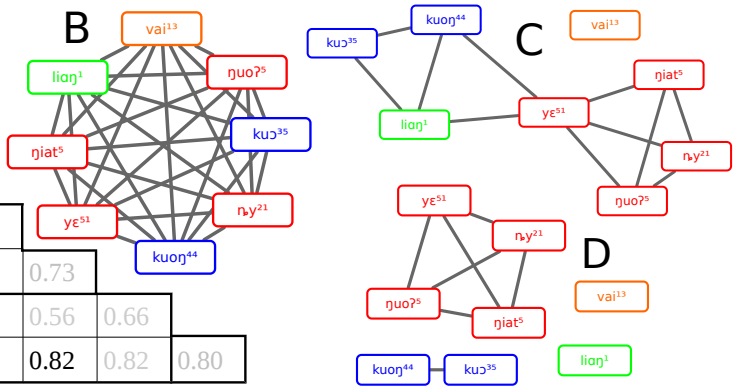

Figure 1: Similarity networks for partial cognate detection. A shows pairwise SCA distances computed between all morphemes of Chinese dialect words for `moon'. Values shaded in gray are excluded following filtering rules 1 and 2 (see text). B shows the initial similarity network with all nodes connected. C shows the network after filtering, and D shows the network after applying the partitioning algorithm.

(2014b). As a general rule, language-specific approaches outperform language-independent ones, provided the sample size is large enough (List 2014a).

Two similarity measures are used in this paper, one language-independent, and one language-specific one. The above-mentioned SCA method for phonetic alignments (List 2012b, 2014b) reduces the phonetic space of sound sequences to 28 sound classes. Based on a scoring function which defines transition scores between the sound classes, phonetic sequences are aligned and similarity and distance scores can be determined. The LexStat approach List (2012a, 2014b) also uses sound classes, but instead of using a pre-defined scoring function, transition scores between sound classes are determined with help of a permutation test. In this test, words drawn from a randomized sample are repeatedly aligned with each other in order to create a distribution of sound transitions for unrelated languages. This distribution is then compared with the actual distribution retrieved from aligned words in the word list, and a language-specific scoring function is created List (2014b). SCA is very fast in computation, but LexStat has a much higher accuracy. Both approaches are freely available as part of the LingPy software package (List and Forkel 2016).

3.2 Sequence Similarity Networks

Sequence similarity networks are tools for exploratory data analysis. In evolutionary biology they are used to study complex evolutionary processes (Méheust et al. 2016, Corel et al. 2016). They represent sequences as nodes and connections between nodes represent similarities which are usually determined from similarity scores exceeding a certain threshold (Alvarez-Ponce et al. 2013). Since evolutionary processes leave specific traces in the network topology, they can be identified by applying techniques for network analysis. In linguistics, sequence similarity networks have been rarely applied (Lopez et al. 2013), although they are applicable, provided that one uses informed measures for phonetic similarity.

For the application of sequence similarity networks it is essential to decide when to draw an edge between two nodes and when not. For the new approach to partial cognate detection, three filtering criteria are applied. (1) No edges are drawn between morphemes which occur in the same word. (2) No morpheme in one word is linked to two morphemes in another word, with the preference given to morpheme pairs with the lowest phonetic distance applying a greedy strategy. (3) Edges are only drawn when the phonetic distance between the morphemes is beyond a certain threshold. The application of the filtering criteria is illustrated in Fig. 1 for the exemplary words shown in Table 1.

3.3 Network Partitioning

Cognate morphemes in a similarity network can be found by partitioning the network into groups. Many algorithms are available for this purpose, as can be seen from evolutionary biology, where homology detection is frequently approached from a network perspective (Vlasblom and Wodak 2009). Three different algorithms were tested for this purpose. A flat version of the UPGMA algorithm for hierarchical clustering (Sokal and Mich-

ener 1958), which terminates when a certain user-defined threshold is reached is originally underlying the LexStat algorithm and was therefore also included in this study. Markov Clustering (van Dongen 2000) uses techniques for matrix multiplication to inflate and expand the edge weights in a given network until weak edges have disappeared and a few clusters of connected nodes remain. Markov Clustering is very popular in biology and was shown to outperform the popular Affinity Propagation algorithm (Frey and Dueck 2007) in the task of homolog detection in biology (Vlasblom and Wodak 2009). As a third method, we follow List et al. (2016b) in testing Infomap (Rosvall and Bergstrom 2008), a method that was originally designed to detect *communities* in complex networks. Communities are groups that share more links with each other than outside the group (Newman and Girvan 2004). Infomap uses random walks to find the best partition of a network into communities. Infomap is not a classical partitioning algorithm, and we do not know of any studies which tested its suitability for the task of homolog detection in evolutionary biology, but according to List et al. (2016b), Infomap shows a better performance than UPGMA in automatic cognate detection.

3.4 Analyses and Evaluation

All methods, be it classical or partial cognate detection, require a user-defined threshold. Since our gold standard data was too small to split it into training and tests sets, we carried out an exhaustive comparison of all methods on different thresholds varying between 0.05 and 0.95 in steps of 0.05. B-cubed scores were chosen as an evaluation measure for cognate detection (Bagga and Baldwin 1998), since they have been shown to yield sensible results (Hauer and Kondrak 2011).

With SCA and LexStat, two classical methods for cognate detection were tested List (2014b), and their underlying models for phonetic similarity (see Sec. 3.1) were used as basis for the partial cognate detection algorithm. All in all, this yielded four different methods: LexStat, LexStat-Partial, SCA, and SCA-Partial. Since our new algorithms yield partial cognates, while LexStat and SCA yield ``complete'' cognates, it is not possible to compare them directly. In order to allow for a direct comparison, partial cognate sets were converted into ``complete'' cognate sets using the above-mentioned strict coding approach proposed by Ben Hamed and Wang (2006): only those words in which *all* morphemes are cognate were assigned to the cognate same set. With a total of three different clustering algorithms (UPGMA, Markov Clustering, and Infomap), we thus carried out twelve tests on complete cognacy (three for each of our four approaches), and six additional tests on pure partial cognate detection, in which we compared the suitability of SCA and LexStat as string similarity measures.

LexStat				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.60	0.9030	0.8743	0.8878
Markov	0.50	0.9123	0.8752	0.8933
Infomap	0.50	0.9131	0.8866	0.8995
SCA				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.45	0.8595	0.8707	0.8648
Markov	0.45	0.8049	0.8097	0.8031
Infomap	0.35	0.8901	0.8573	0.8734
LexStat-Partial Complete Cognacy				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.90	0.9193	0.9638	0.9399
Markov	0.70	0.9275	0.9342	0.9298
Infomap	0.65	0.9453	0.9363	0.9404
SCA-Partial Complete Cognacy				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.60	0.9304	0.9045	0.9172
Markov	0.95	0.8153	0.8949	0.8446
Infomap	0.55	0.9104	0.9366	0.9223
LexStat-Partial Partial Cognacy				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.75	0.8920	0.8820	0.8867
Markov	0.60	0.8858	0.8724	0.8782
Infomap	0.60	0.8876	0.8844	0.8856
SCA-Partial Partial Cognacy				
Cluster-Method	**T**	**P**	**R**	**FS**
UPGMA	0.50	0.8597	0.8509	0.8552
Markov	0.50	0.8074	0.7621	0.7755
Infomap	0.35	0.8676	0.8439	0.8553

Table 3: General performance of the algorithms on all datasets. The table shows for each of the 18 different methods the threshold (T) for which the best B-Cubed F-Score was determined, as well as the B-Cubed precision (P), recall (R), and F-score (FS). The best result in each block is shaded in gray.

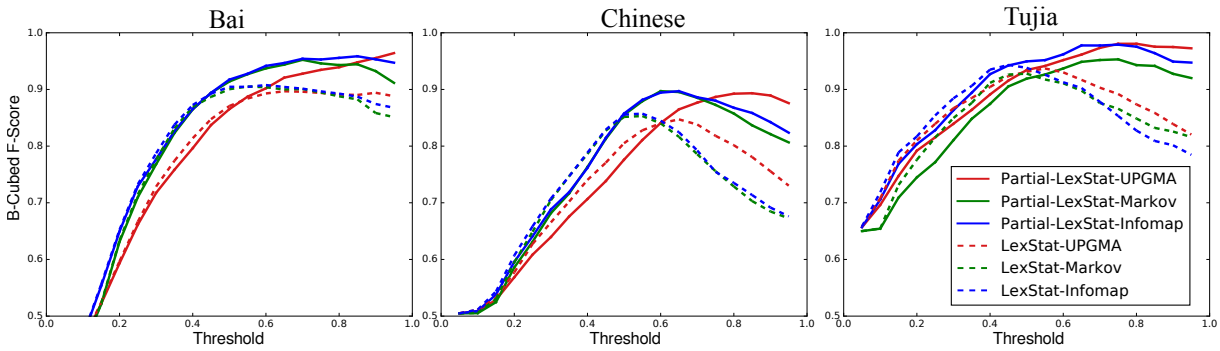

Figure 2: Comparing the results for the LexStat sequences similarities

3.5 Implementation

The code was implemented in Python, as part of the LingPy library (Version 2.5, List and Forkel (2016), `http://lingpy.org`). The Igraph software package (Csárdi and Nepusz 2006) is needed to apply the Infomap algorithm.

4 Results

The aggregated results of the test (thresholds, precision, recall, and F-scores) are given in Table 3, specific results for the comparison of LexStat with LexStat-Partial are given in Table 3. In general, one can clearly see that the partial cognate detection algorithms outperform their non-partial counterparts when applying the complete cognacy measure. The differences are very striking, with LexStat-Partial outperforming its non-partial counterpart by up to four points, and SCA-Partial outperforming the classical SCA variant by almost five points.[2] In contrast, we do not find strong differences in the performance of the cluster algorithms. Infomap outperforms the other cluster algorithms in almost all tests (all other aspects being equal), but the differences are not high enough to make any further conclusions at this point.

When comparing the aggregated results for the true evaluation of partial cognate detection (the last two blocks in Figure 2), the scores are less high than in the complete cognate analyses. Given that we cannot detect any striking tendency, like a drastic drop of precision or recall, this suggests that the algorithms generally loose accuracy in the task of ``true'' partial cognate detection. This is surely not surprising, since the task of detecting exactly which morphemes in the data are historically related is much more complex than the task of detecting which words are completely cognate.

[2]By one point, we mean 0.01 on the B-Cube scale.

In Figure 2, detailed analyses for the LexStat analyses with complete cognate evaluation (the first and the third block in Table 3) are shown for each of the datasets, and throughout all thresholds we tested. The superior performance of the partial cognate detection variants is reflected in all datasets. That the internal diversity of the Chinese languages largely exceeds Bai and Tujia can be seen from the generally lower scores which all algorithms achieve for the datasets.

5 Discussion

This paper has presented a pilot approach for the detection of partial cognates in multilingual word lists. Although the results are very promising at this stage, we can think of many points where improvement is needed, and further studies are needed to fully assess the potential of the current approach. First, it should be tested on additional datasets, and ideally also on language families other than Sino-Tibetan. Second, since our approach is very general, it can easily be adjusted to employ different string similarity measures or different partitioning algorithms, and it would be interesting to see whether alternative measures can improve upon our current version.

Acknowledgments

This research was supported by the DFG research fellowship grant 261553824 *Vertical and lateral aspects of Chinese dialect history* (JML). EB is supported by by the ERC under the European Community's Seventh Framework Programme, FP7/2007-2013 Grant Agreement # 615274.

Supplementary Material

The Sup. Material contains results, benchmark datasets, and code, downloadable at: `https://zenodo.org/record/51328`.

References

David Alvarez-Ponce, Philippe Lopez, Eric Bapteste, and James O. McInerney. 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proceedings of the National Academy of Sciences of the United States of America* 110(17):E1594--1603.

Quentin D. Atkinson and Russell D. Gray. 2006. How old is the Indo-European language family? Illumination or more moths to the flame? In Peter Forster and Colin Renfrew, editors, *Phylogenetic methods and the prehistory of languages*, McDonald Institute for Archaeological Research, Cambridge, pages 91--109.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the ACL*. pages 79--85.

Mahe Ben Hamed and Feng Wang. 2006. Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23:29--60.

Běijīng Dàxué 北京大学, editor. 1964. *Hànyǔ fāngyán cíhuì* 汉语方言词汇[Chinese dialect vocabularies]. Wénzì Gǎigé 文字改革, Běijing 北京.

Eduardo Corel, Philippe Lopez, Raphaël Méheust, and Eric Bapteste. 2016. Network-thinking: Graphs to analyze microbial complexity and evolution. *Trends Microbiology* 24(3):224--237.

Gábor Csárdi and Tamás Nepusz. 2006. The igraph software package for complex network research. *InterJournal Complex Systems* page 1695.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science* 315:972--976.

Leanne S. Haggerty, Pierre-Alain A. Jachiet, William P. Hanage, David A. Fitzpatrick, Philippe Lopez, Mary J. O'Connell, Davide Pisani, Mark Wilkinson, Eric Bapteste, and James O. McInerney. 2014. A pluralistic account of homology: adapting the models to the data. *Mol. Biol. Evol.* 31(3):501--516.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2015. *Glottolog*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of the 5th International Joint NLP conference*. pages 865--873.

Hóu, Jīngyī 侯精一, editor. 2004. *Xiàndài Hànyǔ fāngyán yīnkù* 现代汉语方言音库[Phonological database of Chinese dialects]. Shànghǎi Jiàoyù 上海教育, Shànghǎi 上海.

Gerhard Jäger. 2013. Phylogenetic inference from word lists using weighted alignment with empirical determined weights. *Language Dynamics and Change* 3(2):245--291.

Johann-Mattis List. 2012a. Lexstat. automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS and UNCLH*. Stroudsburg, pages 117--125.

Johann-Mattis List. 2012b. SCA. phonetic alignment based on sound classes. In Marija Slavkovik and Dan Lassiter, editors, *New directions in logic, language, and computation*, Springer, Berlin and Heidelberg, pages 32--51.

Johann-Mattis List. 2014a. Investigating the impact of sample size on cognate detection. *Journal of Language Relationship* 11:91--101.

Johann-Mattis List. 2014b. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf. URL: `http://sequencecomparison.github.io`.

Johann-Mattis List. 2016. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1(2). Published online before print.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016a. *Concepticon: A resource for the linking of concept lists*. Max Planck Institute for the Science of Human History, Jena. Version: 1.0, URL: `http://concepticon.clld.org`.

Johann-Mattis List and Robert Forkel. 2016. *LingPy. A Python library for historical linguistics*. Max Planck Institute for the Science of Human History, Jena. Version 2.5. URL: `http://lingpy.org`. With contributions by Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel, and Simon Greenhill.

Johann-Mattis List, Simon Greenhill, and Russell Gray. 2016b. The potential of automatic cognate

detection for historical linguistics. Manuscript in preparation.

Philippe Lopez, Johann-Mattis List, and Eric Bapteste. 2013. A preliminary case for exploratory networks in biology and linguistics. In Heiner Fangerau, Hans Geisler, Thorsten Halling, and William Martin, editors, *Classification and evolution in biology, linguistics and the history of science*, Franz Steiner Verlag, Stuttgart, pages 181--196.

Wilhelm Meyer-Lübke. 1911. *Romanisches etymologisches Wörterbuch*. Winter, Heidelberg.

Raphaël Méheust, Ehud Zelzion, Debashish Bhattacharya, Philippe Lopez, and Eric Bapteste. 2016. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proceedings of the National Academy of Sciences of the United States of America* 113(3): 3579--3584.

M. E. J. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69(2):026113+.

Martin Rosvall and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America* 105(4):1118--1123.

Damian Satterthwaite-Phillips. 2011. *Phylogenetic inference of the Tibeto-Burman languages*. PhD Thesis, Stanford University, Stanford.

Robert. R. Sokal and Charles. D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* 28:1409--1438.

George S. Starostin. 2013a. Annotated Swadesh wordlists for the Tujia group. In George Starostin, editor, *The Global Lexicostatistical Database*, RGGU, Moscow. URL: `http://starling.rinet.ru/new100/tuj.xls`.

George S. Starostin. 2013b. Lexicostatistics as a basis for language classification. In Heiner Fangerau, Hans Geisler, Thorsten Halling, and William Martin, editors, *Classification and evolution in biology, linguistics and the history of science*, Franz Steiner Verlag, Stuttgart, pages 125--146.

Robert L. Trask. 2000. *The dictionary of historical and comparative linguistics*. Edinburgh University Press, Edinburgh.

Peter Turchin, Ilja Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship* 3:117--126.

Stijn M. van Dongen. 2000. *Graph clustering by flow simulation*. PhD Thesis, University of Utrecht.

James Vlasblom and Shoshana J. Wodak. 2009. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* 10:99.

Feng Wang. 2006. *Comparison of languages in contact*. Academia Sinica, Taipei.