ERC Starting Grant 2016

Annex 1 to the Grant Agreement (Description of the Action) Part B

Action Acronym:	**CALC**
Action Number:	**715618**
Action Title:	**Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics**
Principal investigator:	**Dr. Johann-Mattis List**
Host institution:	**Max Planck Institute for the Science of Human History, Jena, Germany**

Contents

1 Early Achievements Track-Record . . . 3
1 A Personal Statement on my Research . . . 3
2 Refereed Publications in Major Journals and Conference Proceedings . . . 3
3 Research Monographs . . . 4
4 Invited Talks . . . 4
5 Online Resources . . . 4
2 State-of-the-art and objectives . . . 6
1 State-of-the-art . . . 6
2 Objectives . . . 9
3 Methodology . . . 12
1 Computer-Assisted Language Comparison . . . 12
2 A Computer-Assisted Study of the History of the Sino-Tibetan Languages . . . 15
3 Detailed Work Plan . . . 16
4 Collaborations and the Host Institution . . . 17

Section 1: Early Achievements Track-Record

1 A Personal Statement on my Research

The search for challenges is a constant in my academic career. In 2001, I spent one year in Saint Petersburg (Russia), working as a volunteer with Upsala Zirk, an organisation which uses circus arts to prevent children from poor families spending most of their life on the streets. As a juggling instructor, I had regular contact with the kids from the project, and it was due to that contact that I acquired fluency in Russian rather quickly, although I had never studied it before. When I came back to Germany and began to study, I enrolled for Indo-European linguistics, Slavic linguistics, and – Chinese, since I was interested whether there was a language that was more difficult to learn than Russian. I was not disappointed in this regard and from all language courses I took during my undergraduate studies, Chinese remained the most challenging one.

When I finished my Magister studies in 2008, I was a very classical historical linguist, with a background in Indo-European reconstruction, Chinese dialectology, and Old Chinese Phonology. I had devoted my Magister thesis to the comparison of linguistic reconstruction in Indo-European and Chinese linguistics. I had no knowledge about computers, algorithms, and programming languages, but a very strong interest in the methodology of language comparison, and I was not satisfied with the loose way in which methodology is often handled in classical Indo-European and Chinese linguistics. My quest for another challenge brought me to Heinrich Heine University Düsseldorf, where I joined an interdisciplinary project on classification and evolution in biology, linguistics, and the history of science. Had I been spending a lot of time on learning natural languages before, I now turned to artificial languages, especially Python and BASH, but also PhP and C++. It was a completely new world that opened up for me, and I was fascinated by the opportunities it offered. I finally thought I had the means to bring linguistic methodology on formal grounds. I started to develop computer applications for phonetic alignment and cognate detection, and my work culminated in my PhD thesis on 'Sequence comparison in historical linguistics' (published with Düsseldorf University Press in 2014), for which I received the Best Dissertaton Award of the Philosophical Faculty. All programs and scripts I wrote were published as a free Python library for quantitative tasks in historical linguistics (LingPy, `http://lingpy.org`), which I have been constantly developing and updating since then.

At some point, however, when I was working as a post-doc in a research project at Philipps-University Marburg and applied the new algorithms to data from different language families, I realized that the results were not bad but never perfect. I understood that automatic applications cannot replace classical linguists, no matter how smart one tunes the algorithms, since they are designed for big data, but linguists work with small data. A new challenge arose: I was (and I still am) convinced that the methodology of language comparison needs to be formalized. Replacing experts with computers, however, does not solve the problems, since computers cannot handle sparse data. What is needed, is an integrative framework in which the best of the two worlds is combined: the intuition and experience of classical linguists, and the consistency and efficiency of automatic approaches. The new challenge was now to develop the tools that are needed to implement the new framework. To be up to the job, I began to train myself in web-programming, especially the development of graphical user interfaces and the creation of web-based, interactive tools with JavaScript. This challenge, the challenge to develop a new framework for *computer-assisted language comparison*, is an ongoing challenge, but I already know what will be the challenge that follows: to use the framework to enhance the research in comparative linguistics. At the moment, I pursue pilot projects with trained classical historical linguists on Amazonian languages (with Thiago Chacon, University of Brasilia) and Burmish languages (with Nathan W. Hill, SOAS, London), and I hope to expand the framework to Sino-Tibetan studies.

2 Refereed Publications in Major Journals and Conference Proceedings

Research which was carried out independently of my Ph.D. supervisor is marked with an asterisk.

- *List, J.-M., J. Pathmanathan, P. Lopez, and E. Bapteste (2016): **Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics.** *Biology Direct* 11.39. 1-17.

- *List, J.-M., P. Lopez, and E. Bapteste (2016): **Using sequence similarity networks to identify partial cognates in multilingual wordlists.** In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers).* Association of Computational Linguistics. 599-605.
- *List, J.-M. (2016): **Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction.** *Journal of Language Evolution* 1.2. 119-136.
- Chacon, T. and *List, J.-M. (2015): **Improved computational models of sound change shed light on the history of the Tukanoan languages**. *Journal of Language Relationship* 13.3. 177-204.
- *List, J.-M. (2015): **Network perspectives on Chinese dialect history**. *Bulletin of Chinese Linguistics* 8. 36-63.
- List, J.-M., S. Nelson-Sathi, H. Geisler, & W. Martin (2014): **Networks of lexical borrowing and lateral gene transfer in language and genome evolution**. *Bioessays* 36.2. 141-150.
- *List, J.-M. (2014): **Investigating the impact of sample size on cognate detection**. *Journal of Language Relationship* 11. 91-101.
- *List, J.-M. & S. Moran (2013): **An open source toolkit for quantitative historical linguistics**. In: *Proceedings of the ACL 2013 System Demonstrations*. Association for Computational Linguistics. 13-18.
- *List, J.-M. (2012): **SCA. Phonetic alignment based on sound classes**. In: Slavkovik, M. and D. Lassiter (eds.): *New directions in logic, language, and computation*. Springer: Berlin and Heidelberg. 32-51.
- Nelson-Sathi, S., J.-M. List, H. Geisler, H. Fangerau, R. Gray, W. Martin, & T. Dagan (2011): **Networks uncover hidden lexical borrowing in Indo-European language evolution**. *Proceedings of the Royal Society B* 278.1713. 1794-1803.

3 Research Monographs

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

4 Invited Talks

- *List, J.-M. (2016): **CLICS 2016. Chances and challenges.** Paper, presented at the workshop "Lexical Semantic Networks and Language Change" (2016/03/17-18, Santa Fe, Santa Fe Institute).
- *List, J.-M. (2015): **Using network models to analyze Old Chinese rhyme data**. Talk, held at the workshop "Recent Advances in Old Chinese Historical Phonology" (2015/11/05-06, London, School of Oriental and African Studies).
- *List, J.-M. (2015): **Datasets and software tools for computer-assisted language comparison**. Paper, presented at the workshop "Databases in Historical Linguistics" (2015/08/20/21, Santa Fe, Santa Fe Institute).
- *List, J.-M. (2015): **The future of the comparative method**. Paper, presented at the conference "Integrating inferences about our past - New findings and current issues in the peopling of the Pacific and South-East Asia" (2015/06/22/23, Jena, Max Planck Institute for the Science of Human History).
- *List, J.-M. (2013): **Improving phylogeny-based network approaches to investigate the history of the Chinese dialects**. Paper, presented at the conference "LFK Society Young Scholars Symposium" (2013/08/11-13, Seattle).

5 Online Resources

All resources are published under free licences, such as GPL 3.0 for software, and CC BY 4.0 for data, and are freely available online under the links specified below as well as on hosting services like GitHub (`http://github.com`) and Zenodo (`http://zenodo.org`).

5.1 Databases

- *List, J.-M. & J. Prokić (2014): Benchmark Database of Phonetic Alignments. Version 1.0. URL: `http://benchmarks.lingpy.org`.
- *List, J.-M., M. Cysouw & R. Forkel (2016): Concepticon. A Resource for the Linking of Concept Lists. Version: 1.0. URL: `http://concepticon.clld.org`.
- *List, J.-M. (2016): EvoBib – A Bibliographical Database for Historical Linguistics. Version 0.18. URL: `http://bibliography.lingpy.org`.

- *List, J.-M, T. Mayer, A. Terhalle & M. Urban (2014): CLICS: Database of Cross-Linguistic Colexifications. Version 1.0. URL: http://clics.lingpy.org.
- *Winter, B., A. Wedel & J.-M. List (2015): The Language Goldmine. Version 0.1. URL: http://languagegoldmine.com.

5.2 Software

- *List, J.-M. & R. Forkel, with contributions by S. Moran, T. Rama, P. Bouda & J. Dellert (2015): LingPy. A Python library for quantitative tasks in historical linguistics. Version 2.5. URL: http://lingpy.org.
- *List, J.-M. (2016): EDICTOR. A JavaScript Application for the Creation, Curation, and Publicaton of Etymological Wordlist Data . Version: 0.1. URL: http://edictor.digling.org.

Section 2: State-of-the-art and objectives

1 State-of-the-art

1.1 Classical Historical Linguistics and the Comparative Method

The comparative method (Meillet 1925 [1954], Weiss 2014) has successfully elucidated the history of a wide range of language families of varying size and age (Baldi 1990, Campbell & Poser 2008) and external evidence has often confirmed the validity of the findings (McMahon & McMahon 2005: 10-14). Thanks to the comparative method linguists have made ground-breaking insights into language change in general and into the history of many specific language families. The comparative method is not just a simple technique, but rather an *overarching framework* to study language history (Fox 1995, Jarceva 1990, Klimov 1990, Ross & Durie 1996). Despite the number of novel computational approaches proposed in the last two decades, the comparative method is still the standard method by which languages are compared and classified.

The comparative method has an underlying workflow that scholars implicitly follow (see Figure 1, following Ross & Durie 1996). The most crucial part is the identification of *cognate words* ② and regular *sound correspondences* ③. Cognate words are words which descend from a common ancestor form (Trask 2000: 63), such as English *father* and Greek πατήρ 'father' which go back to an earlier form ***ph₂tḗr*** in Proto-Indo-European (Mallory & Adams 2006: 209f, Meier-Brügger 2002: 245). Sound correspondences are regularly recurring correspondences of sounds in cognate words (Lass 1997: 130, Trask 2000: 336): English *f-*, for example, corresponds to Ancient Greek π- [p] (*foot*, πούς; *father*, πατήρ; *fire*, πῦρ). The *iterative character* of the workflow requires repetition in all steps. Iteration is important to address circularity problems: *cognate words* ② can, for example, only be identified with help of regular *sound correspondences* ③, but sound correspondences themselves occur only in cognate words. An iterative procedure circumvents this problem by starting with an initial hypothesis regarding sound correspondences and cognate words which is then constantly revised.

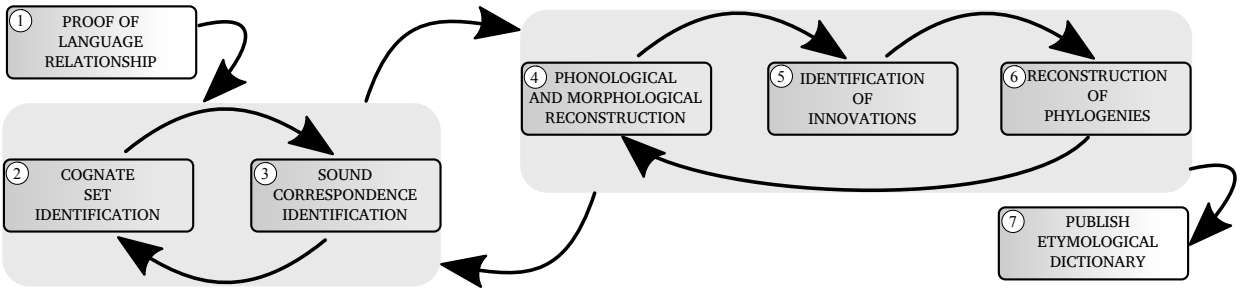


Figure 1: Workflow for the comparative method by Ross and Durie (1996) with two major and multiple minor stages of iteration.

The comparative method is not error-free. Despite its success, it faces a range of serious problems that critics have repeatedly raised: It exhibits a definite lack of *transparency*, in so far as the scholars' intuition plays a major role (Schwink 1994: 29). It shows a lack of *applicability*, because the method is tedious and time consuming (as reflected by the fact that only a fragment of the 7000 languages spoken today have received comparative study). Furthermore, it shows a lack of *reliability* (the degree to which observations are replicable, Liebert & Langenbach Liebert 1995), since neither formal guidelines nor statistical tests are used to arrive at the hypotheses (Baxter & Manaster Ramer 2000: 169-172), which makes it difficult guarantee that scholars working independently will arrive at thesame conclusions (McMahon & McMahon 2005: 26-29).

The comparative method has further weak spots. While linguistic reconstruction of a proto-languages's phonology and lexicon works well on an algebraic level, there is much debate among scholars regarding the concrete realization of reconstructed forms. Already Roman Jakobson claimed that reconstructed languages should confirm to observed tendencies and laws of language typology (Jakobson 1958), but since few have attempted to measure the likelihood of sound transitions (Kümmel 2008), scholars are typically left with their intuition. Similarly, although most scholars now agree that semantic change follows cross-linguistic patterns (Wilkins 1996), and initial attempts of empirical accounts of the probability of semantic shift and the strength of semantic associations are available (Bulakh et al. 2013, **List et al. 2014a**), semantic reconstruction is still based on an ad-hoc identification of potential triggers (Starostin 2010: 100).

1.2 Computational Historical Linguistics

Along with the quantitative turn in historical linguistics in the beginning of the second millenium, many approaches that automate certain parts of the classical workflow of the comparative method are in circulation. The literature is abundant, as are the approaches, and the data to which they are applied. Table 1 contrasts the modules of the classical workflow, as given in Figure 1, with popular automatic approaches. Nearly all of the major modules of the comparative method are addressed in at least one published approach. However, the automatic approaches cannot directly replace the classical comparative method. First, there are *applicability problems*, since the majority of the approaches (preceded by an asterisk in the table) are not publicly available or only suitable for one specific use-case. Secondly, there are *transparency problems*: only a small number of the approaches proposed so far have undergone rigorous testing and evaluation. Usually the approaches operate in a black box fashion that prevents classical linguists from either checking the individual consequences of the inferences, or making use of them to revise their own theories (Prokić & Moran 2013). Last but not least, there are *validity problems*: Quantitative methods are often based on manually pre-compiled datasets (like the collections of cognate sets used for phylogenetic reconstruction, see Atkinson & Gray 2006). Since the quality of the data varies widely, the results often disappoint classical linguists (e.g. Holm 2007), and the methods loose reliability (**Geisler & List 2010**). Scholars also disagree about which evolutionary models to use. Different models may result in widely diverging analyses (compare, e.g., Gray & Atkinson 2003 and Chang et al. 2015).

#	Classical HL	Computational HL	Examples
①	proof of language relationship	probability testing	*Baxter & Manaster Ramer 2000, *Kessler 2001, *Ringe 1992
		phonetic distance	*Jäger 2015
②	cognate set identification	matching sound classes	Turchin et al. 2010
③	sound correspondence identification	phonetic distance and partitioning	**List 2012a, List 2014**, *Steiner et al. 2011
		phonetic alignments	*Kondrak 2000, List 2012b, *Prokić et al. 2009, *Prokić & Cysouw 2013
④	linguistic reconstruction	probabilistic string transducer	*Bouchard-Côté et al. 2013
⑤	identification of innovations	various methods for lexical, gramm., and morphol. data	Chang et al. 2015, Gray & Atkinson 2003, Longobardi et al. 2013, *Ringe et al. 2002
⑥	phylogenetic reconstruction		
⑦	etymologies	(borrowing detection)	*Ark et al. 2007, **List et al. 2014b**, *Nelson-Sathi et al. 2011
		(ancestral state reconstruction)	***Jäger & List 2015, List 2015b**

Table 1: Comparing computational approaches in historical linguistics with the classical comparative method: Approaches in brackets reflect only certain aspects of the original workflow. Examples starred with an asterisk either require specific input data, or their source code has never been published.

Not all computational approaches in historical linguistics are equally popular. The majority of scholars accepts computational approaches to *phylogenetic reconstruction*, be it as a supplement to established classifications, or as an initial heuristic. Phonetic alignment algorithms (**List 2014**, Prokić et al. 2009) are common in dialectology (Wieling & Nerbonne 2015), but less frequent in historical linguistics, mostly because lexical cognacy databases often lack phonetic transcriptions (Dunn 2012, Greenhill et al. 2008). All other computational methods are still in their infancy, although some of the approaches, like those for automatic cognate detection, yield promising results (**List 2014**). Importantly, computational methods have only sporadically addressed the last module of the workflow (step ⑦), the creation of etymological dictionaries which trace 'borrowings, semantic change, and so forth, for the lexicon of the family' (Ross & Durie 1996: 7). This contrasts with the popularity of detailed word histories in classical historical linguistics.

1.3 Sino-Tibetan Historical Linguistics

> '[La] restitution d'une « langue commune » dont le chinois, le tibétain, etc., par exemple, seraient des formes postérieures, se heurte à des obstacles quasi invincibles.' (Antoine Meillet, 1866–1936, 1925 [1954]: 26f)

English translation: 'The reconstruction of a proto-language of which Chinese, Tibetan, etc., are the descendants, encounters almost unsurmountable obstacles.'

With more than 450 identified descendant languages and dialects, the Sino-Tibetan language family is one of the largest language families in the world (Hammarström et al. 2015). Sino-Tibetan languages are spoken across a vast area ranging from Northeast India to South-East Asia (Handel 2008). For comparative historical linguistics and the classical comparative method 'the family presents a complex and challenging picture' (ibid.: 422). Already in 1823, Julius Klaproth (1783–1835) proposed that languages like Chinese, Tibetan, and Burmese

were related (Klaproth 1823, see van Driem 2014). Up to today, however, there is no consensus, regarding the reconstruction of Proto-Sino-Tibetan, or regarding the detailed subgrouping of Sino-Tibetan languages (Handel 2008).

Investigating he history of the Sino-Tibetan languages is particularly hard for three reasons. (1) Language contact is widespread. (2) Sporadic processes of morphological and analogical change mask regular sound change processes. (3) The Sino-Tibetan languages exhibit a high degree of typological diversity. In the Sino-Tibetan language family, language contact is the rule rather than the exception (Thurgood 2003), including contact inside subgroups, among subgroups, or to neighboring language families, like Tai-Kadai, Hmong-Mien, or Austro-Asiatic. Due to the intensive contact within Sinitic, for example, most Chinese dialectologists agree with Norman (2003: 76f) that Chinese is 'not entirely amenable to a Stammbaum formulation'. Due to more than a thousand years of intensive contact between Bai and Chinese languages, Sino-Tibetan linguists disagree whether the Bai languages are the closest relative of Chinese (Starostin 2007, Wang 2006) or a "normal" but heavily siniticized subgroup of Tibeto-Burman (Lee & Sagart 2008, Matisoff 2003).

If language contact can be excluded, sound change is a predominantly regular process that spreads across the whole lexicon of a language (Blevins 2004: 260-268, Kiparsky 1988, Labov 1981). Morphological processes, like suffixation, compounding, or analogy, however, are predominantly sporadic. Morphological processes can mask the regularity of sound change processes and obstruct the identification of regular sound correspondences. Compounding, for example, is a major process of word formation in the Sino-Tibetan family (Matisoff 2003: 153f). If compounds are reduced due to *contraction* (**List 2015a**, Trask 2000: 92), they obscur regular sound correspondences, and this may explain the large-scale inconsistencies in sound correspondences among Sino-Tibetan languages (Handel 2008: 425f). When carrying out a lexical comparison based on word lists, compounding exacerbates the difficulties of identifying cognates, since words across different languages may share only one morpheme which may yield complex patterns of *partial cognacy* (**List 2015b: 56-58**, **List 2016**, Matisoff 2000: 341f, Satterthwaite-Phillips 2011: 99f), as illustrated in Table 2.

Variety	**Form**	**Character**	**Etymological Structure**			
			MC *ŋiot 月	**MC *kwaŋ** 光	**MC *bjut** 佛	**MC *ljaŋH** 亮
Fúzhōu 福州	ŋuoʔ^{5}	月	ŋ u o ʔ ^{5}			
Měixiàn 梅縣	ŋiat^{5} kuoŋ^{44}	月光	ŋ i a t ^{5}	k u o ŋ ^{44}		
Wēnzhōu 溫州	ɳy^{21} kuɔ^{35} vai^{13}	月光佛	ɳ - y - ^{21}	k u ɔ - ^{35}	v a i ^{13}	
Běijīng 北京	yɛ^{51} liɑŋ^{1}	月亮	- y ɛ - ^{51}			l i ɑ ŋ ^{1}

Table 2: Complex etymological structure in word compounds. The table shows partial etymological relations of words for "moon" in four Chinese dialects. Dialect data follows Hóu (2004), Middle Chinese (MC) readings follow Baxter (1992) with modifications.

Sino-Tibetan languages are typologically quite diverse. Tonogenesis (Abramson 2004, Haudricourt 1954), the process by which languages develop tone, occurred frequently and independently in the history of the Sino-Tibetan languages, and sometimes, as in the case of Tibetan, even subgroups have dialects with tone and dialects lacking tone. There are languages with rich inflectional morphology, like the Kiranti languages (Ebert 2003), and languages completely isolating languages, like Chinese (Sun 2006) or Bai (Wiersma 2003). Since words can be easily borrowed, many linguists, including Meillet, see morphology and morphosyntax as stronger evidence for subgrouping than shared vocabulary (Nichols 1996)[1]. However, since many Sino-Tibetan languages lack complex morphology, it is difficult to assemble evidence for deeper affiliations apart from the lexicon.

As a result, proposed subgroupings for the Sino-Tibetan language family differ widely (Handel 2008), as does the evidence scholars use to support their hypotheses (LaPolla 2012). Even the seemingly robust claim that the Sinitic was the first branch to split off (Matisoff 2003, Thurgood 2003) has been challenged on the basis of morphological and lexical evidence (Blench & Post 2013, van Driem 1997) or lack of positive evidence (van Driem 2011, Jacques 2006). On the other hand, new approaches to the reconstruction of Old Chinese, proposed since the late 1980s (Baxter 1992, Starostin 1989, Zhengzhang 2000) and now broadly accepted (Pān 2000, Schuessler 2007), have revolutionized the field. The new reconstructions reveal closer similarities among Old

[1] Longobardi et al. (2013) go even further in arguing that abstract syntax is a stronger indicator of genetic relatedness, although this contradicts the principle by Campbell (2013: 486) that 'permits as evidence of relatedness only comparisons involving sound and meaning together'. Further research will need to show whether Campbell's principle proves right.

Chinese, Tibetan and Burmese (Hill 2014), and come closer to recent reconstructions of Proto-Tibeto-Burman (Matisoff 2003), so that few scholars now doubt that Sino-Tibetan is a valid family (Jacques 2015).

Contrary to the uncertainty in subgrouping and reconstructions, the Sino-Tibetan language family is well represented in terms of data. There are large dialect surveys on the Sinitic languages (Hóu 2004, Běijīng Dàxué 1964) and large collections of lexical data for many Sino-Tibetan varieties spoken in China (Allen 2007, Huáng 1992, Sūn 1991). The STEDT project (Matisoff 2011) makes an invaluable contribution, offering abundant lexical resources on more than 200 Tibeto-Burman languages along with recent Proto-Tibeto-Burman reconstructions (Matisoff 2003). The problem of the available resources is, however, that they are not unified. They differ regarding the questionnaires that scholars used in their field work, regarding the transcription systems employed, and regarding the care with which a given survey was carried out. As a result, the data are largely *unprocessed*: digitally available, but not tagged for meaning, pronunciation, or etymology.

1.4 Summary

The quantitative turn creates a gap between the "new and innovative" quantitative methods and the classical approaches. Classical linguists are often skeptical of the new approaches, partly because the results not necessarily coincide with those achieved by classical methods (Pereltsvaig & Lewis 2015) or only confirm things that linguists already knew (Campbell 2013: 485f), partly because they consider the new approaches which often work in a black box fashion and do not allow one to inspect the concrete findings as overly simplistic (ibid.: 471f). Computational linguists, on the other hand, complain about classical historical linguists' lack of interest in the opportunities which quantitative approaches have to offer, and their lack of consistency when applying the classical methods.

Aspect	**Historical Linguistics**	
	Classical	**Comput.**
inclusion of language-specific knowledge	✓	✗
handling of data sparseness	✓	✗
handling of multiple types of evidence	✓	✗
handling of large amounts of data	✗	✓
consistency	✗	✓
efficiency	✗	✓
quality of results	✓	✗

Table 3: Comparing the advantages and the disadvantages of classical and computational approaches to historical linguistics: While classical approaches are excellent in the inclusion of language-specific knowledge, the handling of data sparseness and multiple types of evidence, computational approaches are superior in consistency, efficiency, and the handling of large amounts of data. Advantages and disadvantages of both disciplines complement each other. Integrating both disciplines would allow us to get the best from both worlds.

Both approaches have their strong and their weak points, as summarized in Table 3. Equipped with deep philological learning, classical linguists have strong intuitions and background knowledge on common and language-specific processes of language change. Basing their analyses on *multiple types of evidence*, classical linguists can work out the most probable solutions even in situations where data is extremely *sparse*. Their disadvantage is that they have difficulties coping with *large amounts of data*. Especially in these situations, the classical comparative method shows a lack of *consistency* and *efficiency*. The advantage of computational linguists is the *efficiency* and *consistency* of their computational models which handle *large amounts of data*. Their disadvantage is that their models can only handle knowledge which has been strictly formalized. They tend to ignore *language-specific aspects* since they can only deal with very *homogeneous types of evidence*. For this reason, computational approaches function poorly with *sparse data*. Since most of the data in historical linguistics *are* sparse, and most of the analyses *need* to reconcile different types of evidence (Sturtevant 1920: 11, Makaev 1977: 88), it is no wonder that the triumphs of computational analyses still lag behind those of classical approaches.

2 Objectives

Computational and classical approaches must complement each other, just as machine translation and human translation complement each other in computer-assisted translation: Current machine translation systems are efficient and consistent, but they are by no means perfect, and no one would use them to replace a trained expert. Trained experts, on the other hand, do not necessarily work consistently and efficiently. In order to enhance both the quality of machine translation and the efficiency and consistency of human translation, *computer-assisted*

translation 'entails an iterative process in which the human translator activity is included in the loop' (Barrachina et al. 2008: 3). Computer-assisted workflows are also common in molecular biology and bioinformatics (Kuiken & Leitner 2000) where interactive tools facilitate manual correction of automated analyses (Jeon et al. 2014).

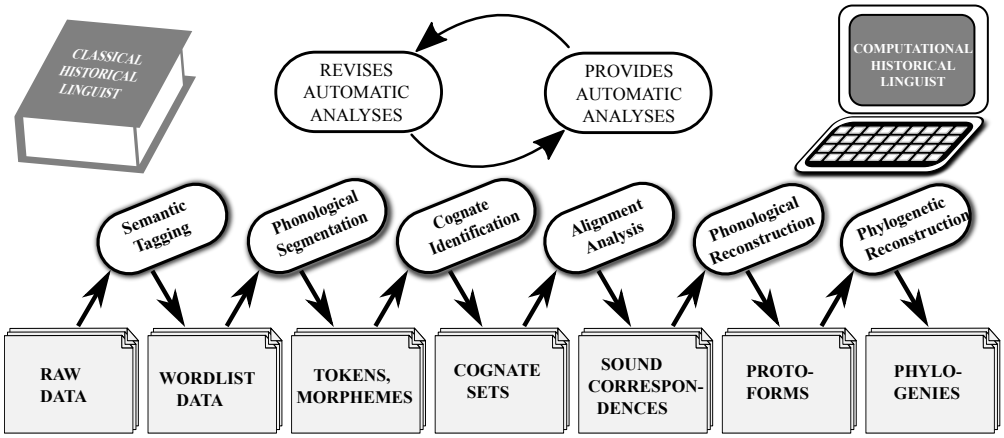


Figure 2: A workflow for *Computer-Assisted Language Comparison*: Data are passed back and forth between a classical and a computational historical linguist. The computational linguists provides initial analyses of the data, and the classical linguist revises them. In this way, one can efficiently get from raw data to proto-forms and phylogenies.

Following the idea of computer-assisted frameworks, a framework for *computer-assisted language comparison* (CALC) becomes the key to reconcile classical and computational approaches in historical linguistics. Computational approaches may still not be able to compete with human experts, but when used to pre-process the data and with human experts systematically correcting the results, they can drastically increase both the efficiency and the consistency of the classical comparative method. Figure 2 shows an exemplary workflow for computer-assisted language comparison which starts from raw data and proceeds in iterative steps up to proto-forms and phylogenies.

The Sino-Tibetan language family is one of the most challenging test cases for both computational and classical methods in historical linguistics. It exposes three major weaknesses in all current approaches: the proper handling of language contact, the complex interaction of lexical and phonological change, and the problem of subgrouping in the absence of morphosyntactic evidence. With the availability of huge collections of digitized resources, too large to manually inspect, the Sino-Tibetan languages are an ideal candidate to test and prove the suitability of the CALC framework. If CALC works on the Sino-Tibetan family, it will also work on other language families.

2.1 Establishing a Framework for Computer-Assisted Language Comparison

The iterative character of the classical comparative method is also a characteristic of computer-assisted language comparison. At all stages, experts can intervene and correct computational results. Three work packages are important for the CALC framework: (1) *software*, (2) *tools*, and (3) *data*. Software refers to transparent automatic methods for the preprocessing of data. Tools refer to web-based interfaces that translate the data in human-readable form and allow for curation and correction by experts. Data refers to gold standards needed to test and train the new automatic methods. The work packages crucially interact: *data* are important for the training and evaluation of the *software*, *software* is used to pre- and post-process the *data*, and *tools* are used to create and curate the *data* by correcting the analyses produced by the *software*.

2.1.1 *Software*

In order to guarantee that experts have free and quick access to the best available methods for the pre- and post-processing of etymological data, the project will create a *unifying software basis* which will (a) *integrate* existing software packages from biology and linguistics in order to provide a basis to test, develop and share *complex workflows*, and (b) offer high quality implementations of new *methods*. In order to *integrate* existing software packages into complex workflows, the project shall write enhanced scripts for the conversion between various data formats. The core *methods* the project will focus on are *improved methods for cognate detection*, new methods for *phonological reconstruction*, and innovative methods for the *reconstruction of word etymologies* (see Section b: 1.1.2).

In order to foster the use of computational methods in historical linguistics and to ease the access for beginners, the project will host a workshop on "Computational Workflows for Quantitative Language Comparison" during its second year. Participants will to contribute to a monograph with the same title. In this monograph researchers and research teams submit tutorials for complex workflows along with their source code, so that readers can directly replicate the examples. The monograph will appear as a hybrid publication which can be purchased in print or downloaded for free.

2.1.2 *Tools and Interfaces*

In order to enable classical historical linguists to inspect, correct, and analyse the outputs produced by automated applications, the project will develop a series of web-based interactive tools and publish them as a free JavaScript library for *digital historical linguistics* along with free sample applications. The core of this library will be a stable and lightweight etymological dictionary editor that allows linguists to create, annotate, analyse, and publish etymological data (see Section b: 1.2.1), along with a suite of JavaScript applications for *interactive data visualization* (see Section b: 1.2.2).

In order to propagate the tools and their usage, a one-week summer school will be launched during the forth year of the project at the host institution in Jena (working title "Computer-Assisted Language Comparison in Practice"). Eight doctoral students who specialize in historical language comparison, including at least four students who work on endangered languages of Asia, will be invited and trained in the framework of computer-assisted language comparison. In addition to the summer school, the project staff will collaboratively set up an exhaustive online manual including exemplary applications in order to enable scholars across the world to make use of the tools in their own research. Furthermore, the principal investigator will write an exhaustive monograph on computer-assisted language comparison (to be published as a hybrid publication with free online access), which will provide a general introduction to the framework of computer-assisted language comparison and serve both as a handbook and a reference.

2.1.3 *Data*

For testing and training of new methods and algorithms, *benchmark databases* are of great importance. Benchmarks or "gold standards" offer authoritative results for tasks in computer science and can be used to test the quality of algorithms (Thompson 2009: 153). A benchmark for the task of *cognate detection in multilingual word lists*, for example, offers human experts' explicit cognate decisions. The suite of benchmark databases will expand the small number of available benchmark datasets (**List 2014, List & Prokić 2014**) and add high quality benchmark datasets for new tasks which will directly feed into the development of the new algorithms (see Section b: 1.1).

2.2 Uncovering the Phylogeny of the Sino-Tibetan Language Family

Digital resources for the Sino-Tibetan language family are abundant, but only a minimal amount of these resources are *processed*. Reconstructions for some branches and subgroups of Sino-Tibetan are available (Mann 1998, Opgenort 2005, VanBik 2009, Wang 2006). There is are also recent reconstructions for Proto-Tibeto-Burman (Matisoff 2003) and Old Chinese (Baxter & Sagart 2014, Pān 2000, Schuessler 2007). There is, however, no accepted reconstruction for the whole family and also no collection of cognate sets assembled across unified word lists. In order to use the available resources for historical comparison, a large number of diverse datasets will be unified during the project. Here, the initial work consists of linking the data to existing resources, unifying the transcriptions systems, and selecting an initial list of comparison concepts which serve as the basis for the compilation of word lists for selected Sino-Tibetan languages (see Section b: 2). The goal is to cover 500 concepts translated into 80 different Sino-Tibetan languages.

The computer-assisted identification of etymologically related words will be carried out in close collaboration with experts in Sino-Tibetan linguistics. Once the cognate identification is finished, The project will apply state-of-the-art methods for phylogenetic reconstruction (Bouckaert et al. 2014) and newly developed methods for phylogenetic tree reconciliation to shed light on the history of the Sino-Tibetan languages. In order to strengthen and coordinate the collaboration with experts, a workshop on "Linguistic reconstruction in Chinese and Sino-Tibetan" will convene during the project's first year. The proceedings of this workshop will appear as a special issue of the *Bulletin of Chinese Linguistics* (Brill, ISSN: 1933-6985).

Section 3: Methodology

1 Computer-Assisted Language Comparison

1.1 Software

1.1.1 *Software Integration and Workflows*

Not all methods relevant for CALC need to be reimplemented from scratch. A large amount of software packages for a variety of tasks are already freely available, not only in biology (Bouckaert et al. 2014, Huson 1998, Than et al. 2008), but also in historical linguistics and dialectology (**List & Moran 2013**, Nerbonne et al. 2011). The problem is to integrate the different software packages, which often only work on specific platforms or require specific input formats. Transparent workflows that can be shared among scholars will facilitate their replication and use. At the moment, even freshly created software packages are seldom made public together with the publications that rely on them (Bouchard-Côté et al. 2013, Downey et al. 2008, Hruschka et al. 2015).

CALC requires unifying software basis which *integrates* existing software packages and provides a basis to test, develop, and share complex workflows. LingPy (`http://lingpy.org`, **List & Moran 2013**), a free[2] Python library for historical linguistics, is built for this very purpose. Apart from advanced methods for phonetic alignment and cognate detection (**List 2014**), it offers also preliminary functionalities to export and import to and from different software packages, and a suite of evaluation routines for testing algorithms with help of benchmark databases (ibid.). In its current form, LingPy offers a solid starting point. During the project, the library will further expand to include enhanced parsers for phylogenetic trees, parsers for NEXUS files (Maddison et al. 1997) which are most frequently used in biological applications, and parsers for the CLDF format specifications (**Forkel et al. 2015**) currently developed by the GlottoBank working group, funded by the host institution.

1.1.2 *New Methods for Computer-Assisted Language Comparison*

Some of the available methods for computational language comparison are sufficiently developed and directly employable in CALC (e.g., methods for phonetic alignments, and phylogenetic reconstruction). Some methods need to be further advanced, especially when dealing with the Sino-Tibetan language family (e.g., methods for cognate detection). Some methods need to be developed from scratch (methods for linguistic reconstruction, and methods for the reconstruction of etymological scenarios). In order to offer sufficient computational support and to take into account the specific challenges of the Sino-Tibetan language family, the project will concentrate on the development of (1) enhanced methods for *cognate and sound correspondence identification*, especially *partial* and *cross-semantic cognate detection* and *borrowing detection*, (2) innovative methods for *phonological reconstruction*, and (3) new methods for the reconstruction of *etymological scenarios* (detailed "word histories"). In order to tackle these problems, three methodological frameworks, which have so far only poorly been studied in computational linguistics, will be used: *sequence similarity networks*, *ancestral state reconstruction*, and *phylogenetic tree reconciliation*. Table 4 shows how the frameworks relate to the three general task and their sub-tasks.

#	General Task	Subtask	Methodolog. Framework
1	enhanced cognate and sound correspondence identification	partial cognate detection	similarity networks
		cross-semantic cognate detection	
		borrowing detection	
2	linguistic reconstruction	phonological reconstruction	ancestral state reconstruction
		lexical reconstruction	
3	reconstructon of etymological scenarios		phylogen. tree reconciliation

Table 4: Relation between the general tasks, the subtasks and the methodological frameworks that the project will use to address the problems.

Sequence Similarity Networks Sequence similarity networks (SSN) are tools for exploratory data analysis. In evolutionary biology, they are used to study complex evolutionary processes like lateral gene transfer (Halary et al. 2013), gene fusion (Jachiet et al. 2013), and gene duplication (Alvarez-Ponce et al. 2013). In SSNs, sequences are represented as nodes and connections between nodes are drawn when the similarity between the sequences exceeds a threshold (ibid.). Since evolutionary processes like lateral transfer or the fusion of

[2] The use of *free* follows the definition for *free software* of the *Free Software Foundation* (`http://www.fsf.org/`).

sequences leave specific traces in the topology of SSNs, they can be identified by applying standard network analysis techniques (Jachiet et al. 2013). In historical linguistics, SSNs have been rarely applied so far (**Lopez et al. 2013**, **List et al. 2016c**), although they are applicable, provided that one uses informed word distance measures which take transition probabilities between sounds into account (**List 2012a, List 2012c**). Especially in the context of Sino-Tibetan language comparison, SSNs may provide great help, since the frequent processes of word compounding, so characteristic for lexical change in the Sino-Tibetan area, yield patterns in SSNs which resemble those of *gene fusion* in biology. In Figure 3 preliminary word similarity networks are reconstructed from multiple alignments of Chinese dialect words (see also **List et al. 2016b**). SSNs in linguistics are not restricted to partial cognate detection: when combining the similarity scores with additional information, like geographic location or known subgroupings of languages, they help to detect recent borrowing events. When combining the scores with information about the strength of semantic associations (**List et al. 2013**), they help to search for cognates across meaning classes in wordlists.

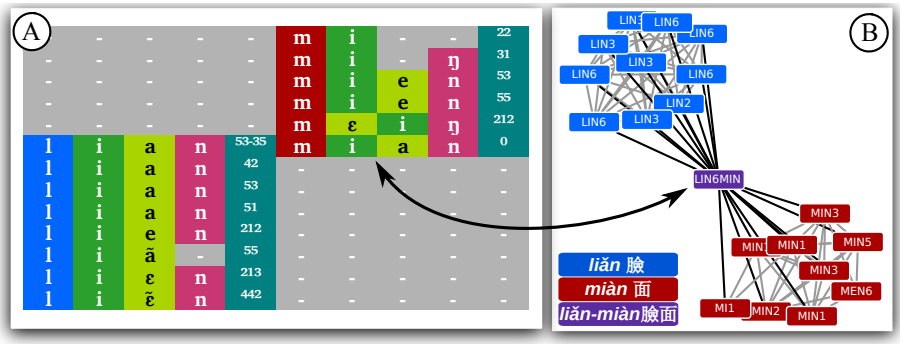


Figure 3: Using word similarity networks for partial cognate detection: The network was reconstructed from pairwise alignments of words for 'face' in 20 Chinese dialects (Hóu 2004). It contains three variants, two simple words of different origin (*liăn* 臉 and *miàn* 面) and one compound in which fuses both words (*liăn-miàn* 臉面). (A) shows a part of the alignment, (B) the resulting SSN in which the fused sequence builds a hub.

Ancestral State Reconstruction Techniques to reconstruct unattested ancestral languages are very common in historical linguistics and have been used for more than 150 years (Schleicher 1861 [1866]). In *phonological reconstruction*, linguists infer proto-sounds and proto-forms which are not reflected in any written source. The inference of ancestral states is also quite common in evolutionary biology, where various techniques for *ancestral states reconstruction* are at hand (Pagel 1999). In computational historical linguistics, however, techniques for phonological reconstruction received scant attention (Bouchard-Côté et al. 2013, Lowe & Mazaudon 1994). The project will develop innovative automatic techniques for phonological reconstruction. In order to make the methods suitable for CALC, several improvements on earlier approaches are necessary. For phonological reconstruction, the most important points include (1) the incorporation of *directionality constraints*, (2) the incorporation of *suprasegmental context constraints*, and (3) the incorporation of *latent states* into the models. Directionality constraints are needed, since many sound changes follow explicit unidirectional pathways. While it is, for example, very probable for a [p] to turn into an [f], the opposite process is very rare (Haspelmath 2004: 19). **Chacon & List (2015)** report promising results for pilot studies on the modeling of directionality constraints in a parsimony framework. The project will further expand the pilot studies on directionality constraints and adapt them to a probabilistic framework (Bouckaert et al. 2014). Suprasegmental context constraints which exceed simple *bigram-models* of context (Bouchard-Côté et al. 2013) are needed in order to capture complex sound change phenomena involving *suprasegmental levels*. This improvement is especially important when working with Sino-Tibetan languages, where supra-segmental sound change processes, like tono-genesis, are very frequent. For example, the tone categories of Middle Chinese, traditionally labelled as *píng* 平, *shăng* 上, *qù* 去, and *rù* 入, result from earlier suffixes and codas (*shăng* < -ʔ, *qù* < -s, *rù* < -[p,t,k], Baxter & Sagart 2014: 194-197). **List (2014: 119-133)** elaborated the foundations for the modeling of the suprasegmental level in sequence comparison, and **Chacon & List** developed them further. In the project these approaches will serve as the starting point for modeling of suprasegmental and segmental aspects of phonetic context in sound change. While the alphabets from which biological sequences are drawn do *not* change, the phoneme systems of languages *change* with time (Garret 2014, **Geisler & List 2013: 121f**), or, as Leonard Bloomfield (1887-1949) put it: '*phonemes change*' (Bloomfield 1933 [1973]: 351). As a result, there is no theoretical justification to assume that an ancestral language has only those sounds which are still reflected in descendent languages. If trained linguists find a [tʃ] corresponding with a [h] in two languages, they immediately reconstruct a [k] for

the ancestral language. They know that a [k] can palatalize to [tʃ] and spirantize to [h] (via [x]). This is what happened with Proto-Indo-European **ḱm̥tóm* 'hundred' (Mallory & Adams 2006: 333f), which became *hundert* [hʊndɐt] in German and *cento* [tʃɛnto] in Italian.[3] All recent approaches ignore that sounds which are not observed in the data may well be reconstructed for ancestral languages (Bouchard-Côté et al. 2013, Hruschka et al. 2015). The project will build on promising pilot studies (**Chacon & List 2015**) to handle *latent character states* in phonological reconstruction.

Phylogenetic Tree Reconciliation While the slogan 'chaque mot a son histoire' has been popular in linguistics for more than 100 years (Campbell 1999: 189), biologists have only recently detected that the history of genes may show striking *mosaic* patterns (Bapteste et al. 2013): the history of a gene family across different species does not necessarily follow the general history of the species. Mosaic patterns result from specific processes which can be incongruent with the species phylogeny, like *gene duplication* or *lateral gene transfer*. Biologists have developed sophisticated models to search and reconcile incongruent patterns. They reconstruct individual trees for homologous genes and project them onto the general species phylogeny. Incongruent patterns are modeled as gene duplication or lateral transfer events. As a result, individual histories of gene families are inferrable with models much more sophisticated than classical birth-death models of gene family evolution (Szöllősi et al. 2015). Given that the process of gene duplication is quite similar to the process of morphological change processes like suffixation or compounding (**List 2014: 41-44**), it is straightforward to employ tree reconciliation techniques in historical linguistics. Figure 4 illustrates how a *word tree*, consisting of compound words for 'moon' in four Chinese dialects, is reconciled with a *language tree*. Phylogenetic tree reconciliation comes quite close to the last module of the workflow of the comparative method (see Figure 1: ⑦). Within a framework of computer-assisted language comparison these techniques will provide invaluable help to speed up the tedious process of etymology reconstructions.

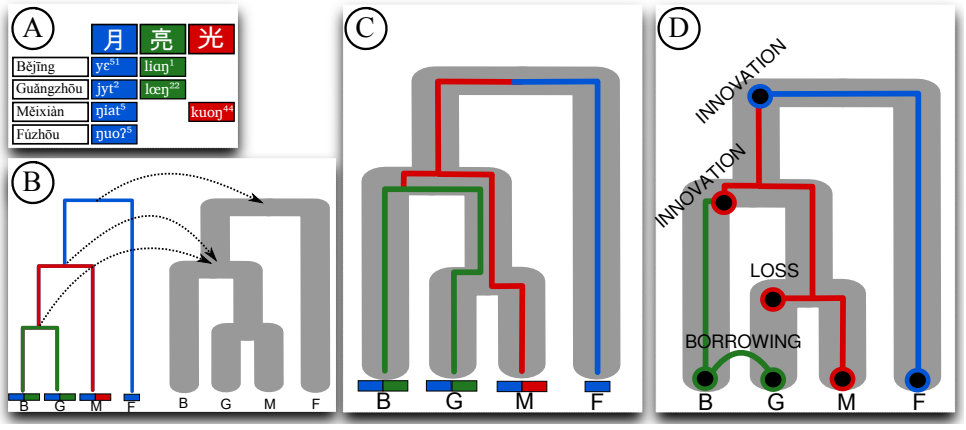


Figure 4: Word tree reconciliation. The Chinese dialects show different compound words for 'moon' Ⓐ. By inferring tree from this pattern Ⓑ and reconciling it with the known phylogeny of the four dialects Ⓒ, one can infer an etymological scenario for the pattern, involving innovations (change from 月光 > 月亮), and even a borrowing event (Guǎngzhōu borrowed 月亮 from Běijīng, Ⓓ).

1.2 Tools and Interfaces

1.2.1 *The etymological dictionary editor*

The etymological dictionary editor will serve as the main interface to exchange data between computational and classical linguists. The PI has already developed a prototype (working title EDICTOR), accessible at `http://edictor.digling.org`. Collaborators from Brasilia (Thiago Chacon), London (Nathan W. Hill) and Paris (Guillaume Jacques) currently test it in different projects. In contrast to existing database systems like RefLex (Segerer & Flavier 2015), or STARLING (Starostin 2000), the etymological dictionary editor is optimized for computer-assisted workflows, allowing for quick data import and export, and offering specific modules for data inspection and correction, including menus for the editing of alignments and cognate sets. The tool features

[3] A famous example of *latent states* in linguistic reconstruction are the *coefficients sonantiques*, later named *laryngeals* (Zgusta 2006), which Ferdinand de Saussure (1857-1913) proposed in 1879. Based on the internal comparison of Greek and Sanskrit morphology, Saussure reconstructed two sounds for Indo-European which were not preserved in any of its descendant languages. 50 years later, after Hrozný (1915) deciphered Hittite, Kuryłowicz (1927) could show that one of the sounds was still reflected in Hittite (compare the initial in Hittite *ḫant-s* 'front, face' with Latin *ante*, Meier-Brügger 2002: 243). Today, scholars agree that at least three laryngeals were present in the sound system of Proto-Indo-European (Clackson 2007: 33-40, Mallory & Adams 2006: 48-50).

various explicit and implicit consistency checks that ensure that edited data passed back to the computational linguist is still machine-readable. The project will enhance and expand the prototype in close collaboration with classical linguists. Further modules include a phoneme analyser linked to existing databases of cross-linguistic phoneme inventories like PBase (Mielke 2008), or PHOIBLE (Moran et al. 2014), and a tool for the inspection of sound correspondences which also allows for user-defined context specifications. In order to guarantee the tool's flexibility, maintainability, accessibility, and availability, the project will distribute it as a free web-based JavaScript application. The advantage of JavaScript applications is that they can run on all platforms and work independently of a server.

1.2.2 *Suite of visualization tools*

For the suite of visualization tools, the project will focus on *interactive visualizations*, including *phylogenetic trees* which visualize the results of phylogenetic tree reconciliation analyses by showing how words evolved along a given reference phylogeny, *networks*, which visualize similarities and differences between etymological data as produced by sequence similarity network approaches, and *geospatial visualizations* which show linguistic variation in geographic space. The PI has created a set of examplary applications which will serve as the starting point for all ongoing work in the project. These applications are available at `http://js.digling.org`. The project will distribute the visualization tools as a free JavaScript library with sample applications and extensive manuals for users and developers (working title *digling.js*).

1.3 Data

So far, only a very small number of benchmark databases for computational historical linguistics are available, including databases for phonetic alignments (**List & Prokić 2014**) and cognate detection (**List 2014: 235**). For the methods to be developed in the project, new benchmark databases are required. The primary target are high-quality benchmarks for cognate detection and phonological reconstruction. In order to assemble the benchmark data, the project will pursue different strategies. Published datasets which give both phonetic transcriptions and cognate judgments (Galucio et al. 2015, Grollemund et al. 2015, Starostin 2013) can be adapted with little effort. The close collaboration with Thiago Chacon (University of Brasilia) who works on large etymological databases of Southern American languages gives us early access to suitable databases for phonological reconstruction and cross-semantic cognate detection. For a few specific tasks, like partial cognate detection, the project will build the benchmark databases from scratch.

2 A Computer-Assisted Study of the History of the Sino-Tibetan Languages

A major challenge in compiling etymological databases is the *harmonization* of resources. As a general strategy for harmonization, the project will start by linking the resources to cross-linguistic databases. Languages and dialects will be linked to Glottolog (Hammarström et al. 2015), and the concept labels in the questionnaires will be linked to the Concepticon (**List et al. 2016a**). The Concepticon is a resource to link the glosses used to denote meanings in different questionnaires. Scholars usually describe the meaning of words with help of short glosses. Since such glosses or *concept labels* were never standardized and meaning is inherently fuzzy, different datasets differ widely in their choice of concept labels. The basic concept GREASE for example is labelled 'animal oil' in Allen (2007), 'fat (grease)' in Swadesh (1955), 'grease (= fat/grease)' in Benedict (1976), 'fat/grease' in Matisoff (1978), 'pig oil' in Ben Hamed & Wang (2006), and 'fat (pig)' in Huáng (1992). Although all the sources *intend* to denote the same concept, they label it in a manner that makes it difficult for both humans and machines to identify words with the same meaning across resources. The Concepticon offers resources to link different questionnaires in a semi-automatic way. An automatic procedure compares questionnaires for rough similarities in concept labels, and an expert later corrects the linking.

This first stage of harmonization guarantees the general comparability of the resources and enables the selection of a first list of *comparison concepts* (similar to *comparative concepts* in typology, Haspelmath 2010), which is important for the initial identification of cognate words. Further steps include the unification of the transcription systems and the "cleaning" of lexical entries: obvious errors in the sources or the digital form of the data need to be corrected; phonetic transcriptions need to be harmonized and represented in such a way that they can be parsed by the software. At all stages, the project will follow the iterative strategy of CALC, according to which computational linguists carry out data pre-processing, and experts correct the results. The

PI has compiled a preliminary database which covers a limited number of comparison concepts in collaboration with G. Jacques and L. Sagart (CRLAO, Paris). This database serves as a proof of concept and can be accessed at `http://sinotibetan.digling.org`.

In order to allow for a consistent handling of the identification of cognate words, the project will follow a strict *bottom-up strategy*: cognates are first identified for smaller subgroups, before cognate sets across groups are identified. The bottom-up strategy minimizes errors in cognate identification, both for humans and machines. Before turning to *cross-semantic cognate identification* cognates will identified inside identical semantic categories in order to make sure that semantic reconstruction is transparent and strict. The project will work in close collaboration with experts from Seattle (Zev Handel, University of Washington), Paris (Laurent Sagart and Guillaume Jacques, both CRLAO), and London (Nathan W. Hill, SOAS) to identify cognate words across all languages within a computer-assisted workflow. In contrast to all previous collections of cognate sets for large language families (Dyen et al. 1992, Dunn 2012, Greenhill 2015, Greenhill et al. 2008, Grollemund et al. 2015), cognate sets will be *aligned*. Aligning the words ensures that all data is represented as transparently as possible and that all regular sound correspondences can be tested, be it with help of automatic methods or by manual inspection. Once the database has been compiled, a large arsenal of methods for phylogenetic reconstruction can be applied to gain deeper insights into the history of the Sino-Tibetan languages, including Bayesian frameworks (Bouckaert et al. 2014), network approaches (**List et al. 2014b**, Nakhleh et al. 2005), and methods for phylogenetic tree reconciliation.

3 Detailed Work Plan

3.1 The Research Team

The research team will consist of the PI, one post-doc with expertise in computational and Sino-Tibetan linguistics, and two doctoral students, one with a background in computational linguistics and data visualization, and one with a background in classical and Sino-Tibetan historical linguistics. A programmer with experience in Python and Javascript and a student assistant with expertise in either computer science or classical historical linguistics will complement the research team. Table 5 shows how labor and responsibilities of the four scientists and the programmer are divided across the four work packages.

	Software	Tools	Data	Sino-Tibetan
Principal Investigator	apply develop	design use	compile pre-process	compile pre-process
Post-Doc	apply develop	design use	compile pre-process	compile pre-process
Classical PhD		use	compile	compile
Computational PhD	apply develop	design	pre-process	pre-process
Programmer	develop	design		

Table 5: How team members contribute to the four work packages.

3.2 Project Goals and Time Schedule

The concrete aims of the project include four primary goals:

`Methods` Develop novel methods and integrate existing methods into a framework for CALC.

`Tools` Create an etymological dictionary editor and a suite of visualization tools for CALC.

`Data` Compile benchmark datasets to assist the development of software for CALC.

`Sino-Tibetan` Use the CALC framework to compile an etymological database of Sino-Tibetan languages.

Each of these major goals can be further subdivided into milestones:

`Methods`
- **Unification:** Unify existing software packages.
- **Partial Cognates:** Develop methods for partial cognate identification.
- **Cross-Semantic Cognates:** Develop methods for cross-semantic cognate identification.
- **Borrowing:** Develop methods for borrowing detection.
- **Ancestral States:** Develop methods for ancestral state reconstruction.

– **Tree Reconciliation:** Develop methods for phylogenetic tree reconciliation.

Tools – **EDICTOR:** Develop an interactive etymological dictionary editor.

– **Visualization Suite:** Develop a suite of interactive visualization tools for CALC.

Data – **Benchmarks:** Compile a suite of benchmark databases for computational historical linguistics.

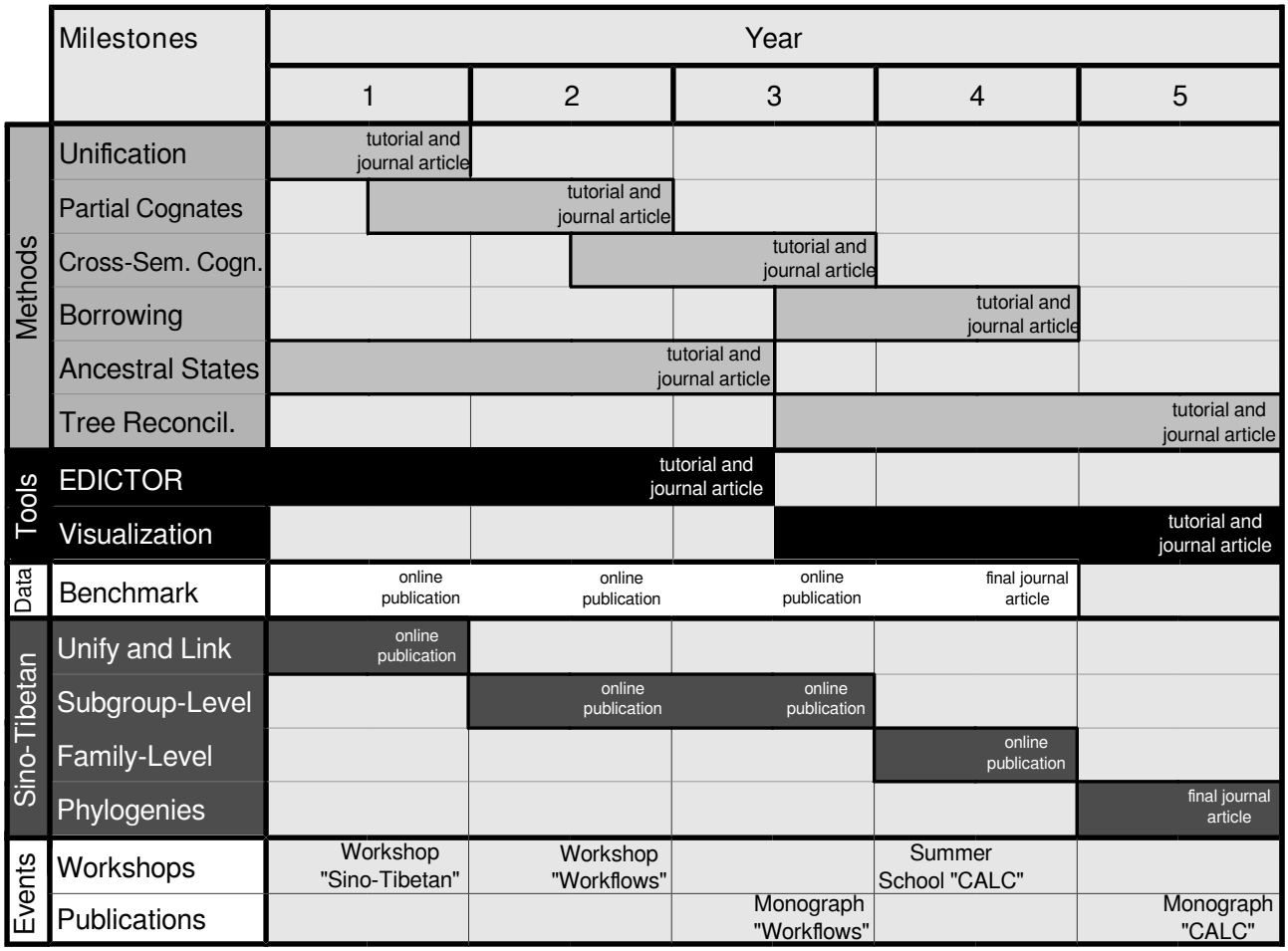
Sinotibetan – **Unify and Link:** Create linked resources of Sino-Tibetan languages with a unified transcription system.

– **Subgroup-Level:** Identify cognates for subgroups and align cognate words.

– **Family-Level:** Identify cognates across subgroups and align cognate words.

– **Phylogenies:** Reconstruct phylogenies and word histories of Sino-Tibetan.

The following time table gives an overview of the schedule of work foreseen, also listing the planned events, including the workshops "Linguistic Reconstruction in Chinese and Sino-Tibetan" and "Computational Workflows for Quantitative Language Comparison", the summer school "Computer-Assisted Language Comparison in Practice", and the publication of "Computational Workflows for Quantitative Language Comparison" (multi-authored volume), and "Computer-Assisted Language Comparison" (written by the PI).



4 Collaborations and the Host Institution

4.1 The Host Institution

The Department of Linguistic and Cultural Evolution at the Max Planck Institute for the Science of Human History (MPI-SHH, Jena, Germany) offers ideal conditions to pursue the interdisciplinary research that is needed in order to develop a new framework for computer-assisted language comparison. In his mission statement, Prof. Russell Gray describes languages as 'documents of history' in which 'a vast amount of information about our past is inscribed'. CALC will increase the efficiency and consistency by which we infer our past from the present of our languages. The interdisciplinary orientation of the MPI-SSH and the close collaboration between

classical and computational linguists offer ideal conditions to develop the new framework. As a member of the GlottoBank Working Group (`http://glottobank.org`), funded by the MPI-SSH under direction of R. Gray, I am in regular contact with researchers and associates of the institute. The current staff of the MPI-SSH consists of excellent programmers and linguists. They are experienced in interdisciplinary work, and the project will highly profit from their advice.

Apart from the MPI-SSH, the project will closely collaborate with researchers from Friedrich-Schiller University Jena. Prof. Martin Kümmel (Chair for Indo-European linguistics) is an expert on Indo-European linguistics and language change, especially sound change. Prof. Volker Gast (Department of English and American Studies) is an expert in corpus linguistics and language typology. The Jena Center for Bioinformatics (`http://www.jcb-jena.de`) offers possibilities for further collaborations.

4.2 Collaborations

I am in regular contact with experts from classical, computational, and Sino-Tibetan historical linguistics, as well as with experts in evolutionary biology. During the project, this collaboration with scholars across disciplines will be further pursued and intensified. The development of new methods for language comparison (see Section b: 1.1), will be carried out in close collaboration with computer scientists and evolutionary biologists. For phonological reconstruction and phylogenetic tree reconciliation, the team of Dr. Remco Bouckaert (Department of Computer Science, University of Auckland) will offer expertise and resources. For the application of similarity networks in historical linguistics, Dr. Eric Bapteste and Dr. Philippe Lopez (team *Adaptation, Intégration, Réticulation, Evolution*, Université Pierre et Marie Curie, Paris) will be close collaboration partners. Dr. Thiago Chacon (Universidade de Brasília), an expert on Southern American languages, is currently testing the first prototype of the etymological dictionary editor and will provide benchmark data from Amazonian language families to test our algorithms for phonological reconstruction.

The development of the etymological database of Sino-Tibetan languages will be carried out in close collaboration with an international team of experts in Sino-Tibetan languages and linguistics. As a former member of the STEDT project and a highly renowned expert in Chinese and Sino-Tibetan linguistics, Prof. Zev Handel (Department for Asian Languages and Literature, University of Washington, Seattle) will provide theoretical advice and practical support. In my role as a post-doctoral research fellow at CRLAO (Paris), I am in regular contact with Dr. Laurent Sagart, an expert on Old Chinese and Chinese dialectology, and Dr. Guillaume Jacques, an expert on Tangut, rGyalrong, and Kiranti languages (both CRLAO, Paris). We have closely collaborated during the past year. L. Sagart and G. Jacques provided assistance in establishing the current prototype of the Sino-Tibetan database. As an external project associate in the ERC synergy project *Beyond Boundaries* (`http://asiabeyondboundaries.org`), I am closely collaborating with one of the principal investigators of the project, Dr. Nathan W. Hill (SOAS, London). He is an established expert in Tibetan and currently works on the reconstruction of Proto-Burmish. In this endeavor, we are currently testing prototypes for the framework of computer-assisted language comparison. With Prof. Dr. Balthasar Bickel (University Zürich), a renowned expert on computational approaches in linguistics and Sino-Tibetan languages will bring in important experience in both phylogenetic approaches and Sino-Tibetan linguistics.

The collaboration between the PI and his team and the external collaborators is of no financial nature, but purely based on common research interests: Apart from travel costs for the invitation of the collaborators or travels of the team to meet the collaborators at their institutions, no further costs will arise for the ERC action.

References

Abramson, A. S. (2004). "The plausibility of phonetic explanations of tonogenesis". In: *From traditional phonology to modern speech processing: Festschrift for Professor Wu Zongji's 95th birthday*. Ed. by G. Fant, H. Fujisaki, J. CAo & Y. Xu. Beijing: Foreign Language Teaching and Research Press, 17–29.

Allen, B. (2007). *Bai Dialect Survey*. SIL International.

Alvarez-Ponce, D., P. Lopez, E. Bapteste & J. O. McInerney (2013). "Gene similarity networks provide tools for understanding eukaryote origins and evolution". *Proc. Natl. Acad. Sci. U.S.A.* 110.17, E1594–1603.

Ark, R. van der, P. Mennecier, J. Nerbonne & F. Manni (2007). "Preliminary identification of language groups and loan words in Central Asia". In: *Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons*. (Borovets, 09/03/2007), 13–20.

Atkinson, Q. D. & R. D. Gray (2006). "How old is the Indo-European language family? Illumination or more moths to the flame?" In: *Phylogenetic methods and the prehistory of languages*. Ed. by P. Forster & C. Renfrew. Cambridge, Oxford, and Oakville: McDonald Institute for Archaeological Research, 91–109.

Baldi, P., ed. (1990). *Linguistic change and reconstruction methodology*. Berlin; New York: Mouton de Gruyter.

Bapteste, E. et al. (2013). "Networks: expanding evolutionary thinking". *TRENDS Genet.* 29.8, 439–441.

Barrachina, S. et al. (2008). "Statistical approaches to computer-assisted translation". *Comput. Linguist.* 35.1, 3–28.

Baxter, W. H. (1992). *A handbook of Old Chinese phonology*. Berlin: de Gruyter.

Baxter, W. H. & A. Manaster Ramer (2000). "Beyond lumping and splitting: Probabilistic issues in historical linguistics". In: *Time depth in historical linguistics*. Ed. by C. Renfrew, A. McMahon & L. Trask. Cambridge: McDonald Institute for Archaeological Research, 167–188.

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

Běijīng Dàxué 北京大學, ed. (1964). *Hànyǔ fāngyán cíhuì* 漢語方言詞匯 [Chinese dialect vocabularies]. Běijing 北京: Wénzì Gǎigé 文字改革.

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

Benedict, P. K. (1976). "Sino-Tibetan: Another Look". *J. Am. Oriental Soc.* 96.2, 167–197.

Blench, R. & M. W. Post (2013). "Rethinking Sino-Tibetan phylogeny from the perspective of North East Indian languages". In: ed. by T. Owen-Smith & N. W. Hill. Berlin and New York: Mouton de Gruyter, 71–104.

Blevins, J. (2004). *Evolutionary phonology. The emergence of sound patterns*. Cambridge: Cambridge University Press.

Bloomfield, L. (1933 [1973]). *Language*. London: Allen & Unwin.

Bouchard-Côté, A., D. Hall, T. L. Griffiths & D. Klein (2013). "Automated reconstruction of ancient languages using probabilistic models of sound change". *Proc. Natl. Acad. Sci. U.S.A.* 110.11, 4224–4229.

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut & A. J. Drummond (2014). "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis". *PLoS Comput. Biol.* 10.4, e1003537.

Bulakh, M., D. Ganenkov, I. Gruntov, T. Maisak, M. Rousseau & A. Zalizniak, eds. (2013). *Database of semantic shifts in the languages of the world*. URL: `http://semshifts.iling-ran.ru/`. (Visited on 11/04/2014).

Campbell, L. (1999). *Historical linguistics. An introduction*. 2nd ed. Edinburgh: Edinburgh Univ. Press.

– (2013). *Historical Linguistics*. 3rd ed. Edinburgh: Edinburgh University Press.

Campbell, L. & W. J. Poser (2008). *Language classification: History and method*. Cambridge: Cambridge University Press.

Chacon, T. C. & J.-M. List (2015). "Improved computational models of sound change shed light on the history of the Tukanoan languages". *J. Lang. Relationship* 13.3, 177–204.

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

Clackson, J. (2007). *Indo-European linguistics*. Cambridge: Cambridge University Press.

Downey, S. S., B. Hallmark, M. P. Cox, P. Norquest & S. Lansing (2008). "Computational feature-sensitive reconstruction of language relationships: developing the ALINE distance for comparative historical linguistic reconstruction". *J. Quant. Linguist.* 15.4, 340–369.

Driem, G. van (1997). "Sino-Bodic". *Bull. Sch. Orient. Afr. Stud.* 60.3, 455–488.

– (2011). "Tibeto-Burman subgroups and historical grammar". *Himalayan Linguist.* 10.1, 31–39.

– (2014). "Tibeto-Burman". In: *The Oxford handbook of Chinese linguistics*. Ed. by W. S.-Y. Wang & C. Sun. Oxford and New York: Oxford University Press, 135–148.

Dunn, M., ed. (2012). *Indo-European lexical cognacy database (IELex)*. URL: `http://ielex.mpi.nl/`.

Durie, M., ed. (1996). *The comparative method reviewed. Regularity and irregularity in language change*. With an intro. by M. D. Ross & M. Durie. New York: Oxford University Press.

Dyen, I., J. B. Kruskal & P. Black (1992). "An Indoeuropean classification. A lexicostatistical experiment". *T. Am. Philos. Soc.* 82.5, iii–132.

Ebert, K. H. (2003). "Kiranti Languages: An overview". In: 505–517.

Forkel, R., M. Dunn, S. Greenhill & J.-M. List (2015). *Cross-linguistic data formats*. GlottoBank Working Group. URL: `http://github.com/glottobank/cldf/`.

Fox, A. (1995). *Linguistic reconstruction. An introduction to theory and method*. Oxford: Oxford University Press.

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

Garret, A. (2014). “Sound change”. In: *The Routledge Handbook of Historical Linguistics*. Ed. by C. Bowern & N. Evans. Routledge, 227–248.

Geisler, H. & J.-M. List (2010). “Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics”. In: *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik*. Ed. by H. Hettrich. Document has been submitted in 2010 and is still waiting for publication. Wiesbaden: Reichert.

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

Gray, R. D. & Q. D. Atkinson (2003). “Language-tree divergence times support the Anatolian theory of Indo-European origin”. *Nature* 426.6965, 435–439.

Greenhill, S. J. (2015). “TransNewGuinea.org: An Online Database of New Guinea Languages”. *PLoS ONE* 10.10, e0141563.

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

Grollemund, R., S. Branford, K. Bostoen, A. Meade, C. Venditti & M. Pagel (2015). “Bantu expansion shows that habitat alters the route and pace of human dispersals”. *Proc. Natl. Acad. Sci. U.S.A.* 112.43, 13296–13301.

Halary, S., J. O. McInerney, P. Lopez & E. Bapteste (2013). “EGN: a wizard for construction of gene and genome similarity networks”. *BMC Evol. Biol.* 13, 146.

Hammarström, H., R. Forkel, M. Haspelmath & S. Bank (2015). *Glottolog*. Version 2.5. URL: `http://glottolog.org`.

Handel, Z. (2008). “What is Sino-Tibetan? Snapshot of a Field and a Language Family in Flux”. *Lang. Linguist. Compass* 2.3, 422–441.

Haspelmath, M. (2004). “On directionality in language change with particular reference to grammaticalization”. In: *Up and down the cline – The nature of grammaticalization*. Ed. by O. Fischer, M. Norde & H. Perridon. Typological Studies in Language. John Benjamins Publishing Company, 17–44.

– (2010). “Comparative concepts and descriptive categories”. *Language* 86.3, 663–687.

Haudricourt, A.-G. (1954). “De l’origine des tons en Vietnamien”. *J. Asiatique* 242, 69–82.

Hill, N. W. (2014). “Cognates of Old Chinese *-n, *-r, and *-j in Tibetan and Burmese”. *Cah. Linguistique – Asie Orientale*, 91–109.

Holm, H. J. (2007). “The new arboretum of Indo-European “trees”. Can new algorithms reveal the phylogeny and even prehistory of Indo-European?” *J. Quant. Linguist.* 14.2-3, 167–214.

Hóu Jīngyī 侯精一, ed. (2004). *Xiàndài Hànyǔ fāngyán yīnkù* 現代漢語方言音庫 [Phonological database of Chinese dialects].

Hrozný, B. (1915). “Die Lösung des hethitischen Problems *[The solution of the Hittite problem]*”. *Mitt. Dtsch. Orient-Ges.* 56, 17–50.

Hruschka, D. J., S. Branford, E. D. Smith, J. Wilkins, A. Meade, M. Pagel & T. Bhattacharya (2015). “Detecting regular sound changes in linguistics as events of concerted evolution”. *Curr. Biol.* 25.1, 1–9.

Huson, D. H. (1998). “SplitsTree: analyzing and visualizing evolutionary data”. *Bioinformatics* 14.1, 68–73.

Huáng Bùfán 黃布凡, ed. (1992). *Zàngmiǎn yǔzú yǔyán cíhuì* 藏緬語族語言詞匯 [A Tibeto-Burman lexicon]. Běijīng 北京: Zhōngyāng Mínzú Dàxué 中央民族大學 [Central Institute of Minorities].

Jachiet, P. A., R. Pogorelcnik, A. Berry, P. Lopez & E. Bapteste (2013). “MosaicFinder: identification of fused gene families in sequence similarity networks”. *Bioinformatics* 29.7, 837–844.

Jacques, G. (2006). *La morphologie du sino-tibétain* [The morphology of Sino-Tibetan]. Paper, presented at the conference “La linguistique comparative en France aujourd’hui” (Paris, 03/04/2006).

– (2015). “The genetic position of Chinese”. In: *Encyclopedia of Chinese Language and Linguistics*. Ed. by R. Sybesma. Leiden and Boston: Brill Online.

Jakobson, R. (1958). “Typological studies and their contribution to historical comparative linguistics”. In: *Proceedings of the Eighth International Congress of Linguistics*, 17–35.

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

Jeon, Y. S., K. Lee, S. C. Park, B. S. Kim, Y. J. Cho, S. M. Ha & J. Chun (2014). “EzEditor: a versatile sequence alignment editor for both rRNA- and protein-coding genes”. *Int. J. Syst. Evol. Microbiol.* 64.Pt 2, 689–691.

Jäger, G. (2015). “Support for linguistic macrofamilies from weighted alignment”. *Proc. Natl. Acad. Sci. U.S.A.* 112.41, 12752–12757.

Jäger, G. & J.-M. List (2015). *Investing the potential of ancestral state reconstruction algorithms in historical linguistics*. Paper, presented at the workshop “Capturing Phylogenetic Algorithms for Linguistics” (Leiden, 10/26–10/30/2015).

Kessler, B. (2001). *The significance of word lists. Statistical tests for investigating historical connections between languages*. Stanford: CSLI Publications.

Kiparsky, P. (1988). “Phonological change”. In: *Linguistics. The Cambridge survey*. Vol. 1: *Linguistic theory. Foundations*. Ed. by F. J. Newmeyer. Cambridge et al.: Cambridge University Press, 363–415.

Klaproth, J. H. (1823). *Asia Polyglotta*. Paris: A. Schubart.

Klimov, G. A. (1990). *Osnovy lingvističeskoj komparativistiki* [Foundations of comparative linguistics]. Moscow: Nauka.

Kondrak, G. (2000). “A new algorithm for the alignment of phonetic sequences”. In: *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. (Seattle, 04/29–05/03/2000), 288–295.

Kuiken, C. L. & T. Leitner (2000). “HIV-1 subtyping”. In: *Computational and evolutionary analysis of HIV molecular sequences*. Ed. by A. G. Rodrigo & G. H. J. Learn. Boston, Dordrecht, London: Kluwer Academic Publishers, 27–54.

Kuryłowicz, J. (1927). “ə indo-européen et ḫ hittite Indo-European ə and Hittite ḫ”. In: *Symbolae grammaticae in honorem Ioannis Rozwadowski*. Ed. by W. Taszycki & W. Doroszewski. Vol. 1. Cracow: Gebethner & Wolf, 95–104.

Kümmel, M. J. (2008). *Konsonantenwandel* [Consonant change]. Wiesbaden: Reichert.

Labov, W. (1981). “Resolving the Neogrammarian Controversy”. *Language* 57.2, 267–308.

LaPolla, R. J. (2012). "Comments on methodology and evidence in Sino-Tibetan comparative linguistics". *Lang. Linguist.* 13.1, 117–132.

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

Lee, Y.-J. & L. Sagart (2008). "No limits to borrowing: The case of Bai and Chinese". *Diachronica* 25.3, 357–385.

Liebert, R. M. & L. Langenbach Liebert (1995). *Science and behavior. An introduction to methods of psychological research*. Englewood Cliffs: Prentice Hall.

List, J.-M. (2012a). "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*. "LINGVIS & UNCLH 2012" (Avignon, 04/23–04/24/2012), 117–125.

– (2012b). "Multiple sequence alignment in historical linguistics. A sound class based approach". In: *Proceedings of ConSOLE XIX*. "The 19th Conference of the Student Organization of Linguistics in Europe" (Groningen, 01/05–01/08/2011). Ed. by E. Boone, K. Linke & M. Schulpen, 241–260.

– (2012c). "SCA. Phonetic alignment based on sound classes". In: *New directions in logic, language, and computation*. Ed. by M. Slavkovik & D. Lassiter. Berlin and Heidelberg: Springer, 32–51.

– (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

– (2015a). "Contraction". In: *Encyclopedia of Chinese language and linguistics*. Ed. by R. Sybesma. Leiden and Boston: Brill Online.

– (2015b). "Network perspectives on Chinese dialect history". *Bull. Chin. Linguist.* 8, 42–67.

– (2016). "Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction". *Journal of Language Evolution* 1.2, 119–136.

List, J.-M. & S. Moran (2013). "An open source toolkit for quantitative historical linguistics". In: *Proceedings of the ACL 2013 System Demonstrations*. "ACL 2013" (Sofia, 08/04–08/09/2013), 13–18.

List, J.-M. & J. Prokić (2014). "A benchmark database of phonetic alignments in historical linguistics and dialectology." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. "LREC" (Reykjavik, 05/26–05/31/2014). Ed. by N. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis. European Language Resources Association (ELRA), 288–294.

List, J.-M., A. Terhalle & M. Urban (2013). "Using network approaches to enhance the analysis of cross-linguistic polysemies". In: *Proceedings of the 10th International Conference on Computational Semantics – Short Papers*. "IWCS 2013" (Potsdam, 03/19–03/22/2013), 347–353.

List, J.-M., T. Mayer, A. Terhalle & M. Urban, eds. (2014a). *CLICS: Database of Cross-Linguistic Colexifications*. Version 1.0. URL: `http://clics.lingpy.org`.

List, J.-M., S. Nelson-Sathi, H. Geisler & W. Martin (2014b). "Networks of lexical borrowing and lateral gene transfer in language and genome evolution". *Bioessays* 36.2, 141–150.

List, J.-M., M. Cysouw & R. Forkel (2016a). "Concepticon. A resource for the linking of concept lists". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. "LREC 2016" (Portorož, 05/23–05/28/2016). Ed. by N. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis. European Language Resources Association (ELRA), 2393–2400.

List, J.-M., J. S. Pathmanathan, P. Lopez & E. Bapteste (2016b). "Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics". *Biol. Direct* 11.39, 1–17.

List, J.-M., P. Lopez & E. Bapteste (2016c). "Using sequence similarity networks to identify partial cognates in multilingual wordlists". In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*, 599–605.

Longobardi, G., C. Guardiano, G. Silvestri, A. Boattini & A. Ceolin (2013). "Toward a syntactic phylogeny of modern Indo-European languages". *J. Hist. Linguist.* 3.1, 122–152.

Lopez, P., J.-M. List & E. Bapteste (2013). "A preliminary case for exploratory networks in biology and linguistics: the phonetic network of Chinese words as a case-study". In: *Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization.* Ed. by H. Fangerau, H. Geisler, T. Halling & W. Martin. Stuttgart: Franz Steiner Verlag, 181–196.

Lowe, J. B. & M. Mazaudon (1994). "The reconstruction engine. A computer implementation of the comparative method". *Comput. Linguist.* 20.3, 381–417.

Maddison, D. R., D. L. Swofford & W. P. Maddison (1997). "NEXUS: an extensible file format for systematic information". *Syst. Biol.* 46.4, 590–621.

Makaev, E. A. (1977). *Obščaja teorija sravnitel'nogo jazykoznanija* [Common theory of comparative linguistics]. Moscow: Nauka.

Mallory, J. P. & D. Q. Adams (2006). *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world*. Oxford: Oxford University Press.

Mann, N. W. (1998). "A phonological reconstruction of Proto Northern Burmic". PhD. Arlington: The University of Texas.

Matisoff, J. A. (1978). *Variational semantics in Tibeto-Burman. The 'organic' approach to linguistic comparison*. Institute for the Study of Human Issues.

– (2000). "On the uselessness of glottochronology for the subgrouping of Tibeto-Burman". In: *Time depth in historical linguistics*. Ed. by C. Renfrew, A. McMahon & L. Trask. Cambridge: McDonald Institute for Archaeological Research, 333–371.

– ed. (2003). *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. University Presses of California, Columbia and Princeton.

– ed. (2011). *STEDT. The Sino-Tibetan Etymological Dictionary and Thesaurus*. University of California at Berkeley. URL: `http://stedt.berkeley.edu/`.

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

Meier-Brügger, M. (2002). *Indogermanische Sprachwissenschaft*. In collab. with M. Fritz & M. Mayrhofer. 8th ed. Berlin and New York: de Gruyter.

Meillet, A. (1925 [1954]). *La méthode comparative en linguistique historique* [The comparative method in historical linguistics]. Repr. Paris: Honoré Champion.

Mielke, J. (2008). *The emergence of distinctive features*. Oxford: Oxford University Press. URL: http://aix1.uottawa.ca/~jmielke/pbase/.

Moran, S., D. McCloy & R. Wright, eds. (2014). *PHOIBLE Online*. URL: http://phoible.org/.

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

Nelson-Sathi, S., J.-M. List, H. Geisler, H. Fangerau, R. D. Gray, W. Martin & T. Dagan (2011). "Networks uncover hidden lexical borrowing in Indo-European language evolution". *Proc. R. Soc. London, Ser. B* 278.1713, 1794–1803.

Nerbonne, J., R. Colen, C. Gooskens, P. Kleiweg & T. Leinonen (2011). "Gabmap – A web application for dialectology". *Dialectologia* Special Issue II, 65–89. URL: http://www.gabmap.nl/.

Nichols, J. (1996). "The comparative method as heuristic". In: *The comparative method reviewed. Regularity and irregularity in language change*. Ed. by M. Durie. With an intro. by M. D. Ross & M. Durie. New York: Oxford University Press, 39–71.

Norman, J. (2003). "The Chinese dialects. Phonology". In: *The Sino-Tibetan languages*. Ed. by G. Thurgood & R. J. LaPolla. London and New York: Routledge, 72–83.

Opgenort, J. R. (2005). *A grammar of Jero. With a historical comparative study of the Kiranti languages*. Leiden: Brill.

Pagel, M. D. (1999). "Inferring the historical patterns of biological evolution". *Nature* 401, 877–884.

Pereltsvaig, A. & M. W. Lewis (2015). *The Indo-European Controversy. Facts and fallacies in historical linguistics*. Cambridge: Cambridge University Press.

Prokić, J. & S. Moran (2013). "Black box approaches to genealogical classification and their shortcomings". In: *Approaches to Measuring Linguistic Differences*. Ed. by L. Borin & A. Saxena. Berlin: Mouton de Gruyter, 437–457.

Prokić, J., M. Wieling & J. Nerbonne (2009). "Multiple sequence alignments in linguistics". In: *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. "LaTeCH-SHELT&R 2009" (Athens, 03/30/2009), 18–25.

Prokić, J. & M. Cysouw (2013). "Combining regular sound correspondences with geographic spread". *Lang. Dyn. Change* 3.2, 147–168.

Pān Wùyún 潘悟雲 (2000). *Hànyǔ lìshǐ yīnyùnxué* 漢語歷史音韻學 [Chinese historical phonology]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.

Renfrew, C., A. McMahon & L. Trask, eds. (2000). *Time depth in historical linguistics*. Cambridge: McDonald Institute for Archaeological Research.

Ringe, D., T. Warnow & A. Taylor (2002). "Indo-European and computational cladistics". *T. Philol. Soc.* 100.1, 59–129.

Ringe, D. A. (1992). "On calculating the factor of chance in language comparison". *T. Am. Philos. Soc.* New Series 82.1, 1–110.

Ross, M. & M. Durie (1996). "Introduction". In: *The comparative method reviewed. Regularity and irregularity in language change*. Ed. by M. Durie. With an intro. by M. D. Ross & M. Durie. New York: Oxford University Press, 3–38.

Satterthwaite-Phillips, D. (2011). "Phylogenetic inference of the Tibeto-Burman languages or on the usefuseful of lexicostatistics (and "megalo"-comparison) for the subgrouping of Tibeto-Burman". PhD thesis. Stanford: Stanford University.

Saussure, F. d. (1879). *Mémoire sur le système primitif des voyelles dans les langues indo- européennes*. Leipzig: Teubner.

Schleicher, A. (1861 [1866]). *Compendium der vergleichenden Grammatik der indogermanischen Sprache*. Vol. 1: *Kurzer Abriss einer Lautlehre der indogermanischen Ursprache*. 2nd ed. Weimar: Böhlau.

Schuessler, A., comp. (2007). *ABC Etymological dictionary of Old Chinese*. Honolulu: University of Hawai'i Press.

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

Segerer, G. & S. Flavier (2015). *RefLex: Reference Lexicon of Africa*. Version 1.1. URL: http://reflex.cnrs.fr.

Starostin, G. S. (2010). "Preliminary lexicostatistics as a basis for language classification: A new approach". *J. Lang. Relationship* 3, 79–116.

– (2013). *Annotated Swadesh wordlists for the Tujia group (Sino-Tibetan family)*. Ed. by G. Starostin. URL: http://starling.rinet.ru/new100/tuj.xls.

Starostin, S. A. (1989). *Rekonstrukcija drevnekitajskoj fonologičeskoj sistemy (Reconstruction of the phonological system of Old Chinese)*. Moscow: Nauka.

– (2000). *The STARLING database program*. URL: http://starling.rinet.ru.

– (2007). "The historical position of Bai". In: *S. A. Starostin: Trudy po jazykoznaniju*. Moscow: Languages of Slavic Cultures, 580–590.

Steiner, L., P. F. Stadler & M. Cysouw (2011). "A pipeline for computational historical linguistics". *Lang. Dyn. Change* 1.1, 89–127.

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

Sun, C. (2006). *Chinese: A linguistic introduction*. Cambridge: Cambridge University Press.

Swadesh, M. (1955). "Towards greater accuracy in lexicostatistic dating". *Int. J. Am. Linguist.* 21.2, 121–137.

Szöllősi, G. J., E. Tannier, V. Daubin & B. Boussau (2015). "The inference of gene trees with species trees". *Syst. Biol.* 64.1, e42–e62.

Sūn Hóngkāi 孫宏開, ed. (1991). *Zàngmiǎnyǔ yǔyīn hé cíhuì* 藏緬語語音和詞匯 [Tibeto-Burman phonology and lexicon]. Zhōngguó Shèhuì Kēxué 中國社會科學 [Chinese Social Sciences Press].

Than, C., D. Ruths & L. Nakhleh (2008). "PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships". *BMC Bioinformatics* 9, 322.

Thompson, J. D. (2009). "Constructing alignment benchmarks". In: *Sequence alignment. Methods, models, concepts, and strategies*. Ed. by M. S. Rosenberg. Berkeley, Los Angeles, and London: University of California Press, 151–177.

Thurgood, G. (2003). "A subgrouping of the Sino-Tibetan languages: The interaction between language contact, change, and inheritance. Phonology". In: *The Sino-Tibetan languages*. Ed. by G. Thurgood & R. J. LaPolla. London and New York: Routledge, 3–21.

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

Turchin, P., I. Peiros & M. Gell-Mann (2010). "Analyzing genetic connections between languages by matching consonant classes". *J. Lang. Relationship* 3, 117–126.

VanBik, K. (2009). *Proto-Kuki-Chin. A reconstructed ancestor of the Kuki-Chin languages*. Berkeley: University of California, Berkeley.

Wang, F. (2006). *Comparison of languages in contact. The distillation method and the case of Bai*. Taipei: Institute of Linguistics Academia Sinica.

Weiss, M. (2014). "The comparative method". In: *The Routledge Handbook of Historical Linguistics*. Ed. by C. Bowern & N. Evans. New York: Routledge. Chap. chapter4, 127–145.

Wieling, M. & J. Nerbonne (2015). "Advances in dialectometry". *Annu. Rev. Linguist.* 1.3, 3–22.

Wiersma, G. (2003). "Yunnan Bai. Phonology". In: *The Sino-Tibetan languages*. Ed. by G. Thurgood & R. J. LaPolla. London and New York: Routledge, 649–673.

Wilkins, D. P. (1996). "Natural tendencies of semantic change and the search for cognates". In: *The comparative method reviewed. Regularity and irregularity in language change*. Ed. by M. Durie. With an intro. by M. D. Ross & M. Durie. New York: Oxford University Press, 264–304.

Zgusta, L. (2006). "The laryngeal and glottalic theories". In: *History of the language sciences. An international handbook on the evolution of the study of language from the beginnings to the present*. Ed. by S. Auroux, E. F. K. Koerner, H.-J. Niederehe & K. Versteegh. Vol. 3. Berlin and New York: de Gruyter, 2462–2479.

Zhengzhang, S. (2000). *The phonological system of Old Chinese*. Trans. by L. Sagart. Paris: École des Hautes Études en Sciences Sociales.