Computer-Assisted Language Comparison

Johann-Mattis List (University of Passau)

1 The quantitative turn in historical linguistics

Background

In the early 1950s, Morris Swadesh (1909–1967) presented a method to measure the genetic closeness between languages on the basis of a statistical formula that was ultimately based on counting the amount of shared cognates across standardized wordlists of different languages (Swadesh 1950). Although it seemed at first that the methods could revive the discipline of historical linguistics, which had past its prime after the structuralist turn in the begin of the 1920s , and had not seen any major methodological or analytical improvement since the begin of the 20th century.[1] Unfortunately, the original interest in the new ideas did not last long, and soon after it was first published, the new method was heavily criticized (Bergsland and Vogt 1962), and went out of vogue some 10 years later.

In the begin of the second millennium, Gray and Atkinson (2003) used similar data but different statistical methods to date the age of the Indo-European language family. They caused a similar stir as Swadesh had done almost half a century ago. But while Swadesh's method was filed away soon after it had been proposed, the method of Gray and Atkinson was part of a general *quantitative turn in historical linguistics*, which started at the begin of the second millennium. This quantitative turn is reflected in a large bunch of literature on such different topics as phonetic alignment (Kondrak 2000, Prokić et al. 2009), automated cognate detection (List 2014), and phylogenetic reconstruction (Atkinson and Gray 2006).

What may have been the reasons why Swadesh's approach was abandoned so quickly by historical linguists?

New studies on language evolution

We can distinguish four different aspects of research approaches in the course of the quantitative turn. As a first and most prominent aspect, we have research dealing with questions of *phylogenetic reconstruction* which usually involved *dating* as well. Language data are not only analyzed to yield a topology of the branching structure of the language family in question, but in addition, absolute branch lengths are often also inferred, which allow to estimate when a given language family has originated. The software and methods used for these studies are usually taken or inspired from approaches developed first in evolutionary biology. As of now, quite a few different language families have been analyzed in this way, including Indo-European (Chang et al. 2015, Gray and Atkinson 2003), Austronesian (Gray et al. 2009), Dravidian (Kolipakam et al. 2018), Bantu (Grollemund et al. 2015), Pama-Nyungan (Bowern et al. 2011), Japonic (Lee and Hasegawa 2011), and Sino-Tibetan (Sagart et al. 2019). In addition, scholars have also attempted to provide unified methods that could be applied in a completely automated fashion to all languages of the world (Holman et al. 2011).

Another strand of research deals with the computation of inference procedures which were traditionally only carried out manually. Most prominently, we find here various attempts to automate different aspects of the general workflow of the traditional *comparative method* for historical language comparison (Weiss 2015). Breaking down the workflow into some of its major parts, we thus find

[1] The last major improvement, the decipherment of Hittite, which also helped to proof that it was an Indo-European language dated back to Hrozný (1915).

(1) automated methods for the comparison of words, as reflected in methods for phonetic alignment (Kondrak 2000, Prokić et al. 2009) and automated cognate detection (Hauer and Kondrak 2011, List et al. 2016, Turchin et al. 2010), (2) automated approaches for the detection of borrowings (List 2015, Mennecier et al. 2016, Nelson-Sathi et al. 2011),[2] (3) automated approaches for linguistic reconstruction (Bouchard-Côté et al. 2013, Jäger 2019), and (4) automated approaches for the detection of sound correspondences (List 2019b).

While the second strand deals mostly with questions of inference, a third strand organizes inferred data in form of large-scale online databases that aggregate different kinds of information on the world's languages. The most prominent of these databases is beyond doubt the *World Atlas of Language Structures* (Dryer and Haspelmath 2013), but in addition we also find attempts to aggregate cross-linguistic information on phoneme inventories (Maddieson et al. 2013, Moran and McCloy 2019), polysemies (List et al. 2018), phonotactics (Donohue et al. 2013), borrowings (Haspelmath and Tadmor 2009), as well as datasets like D-Place, that compare cultural, environmental, and linguistic diversity (Kirby et al. 2016).

While the popular phylogenetic approaches deal with c-linguistics (or p-linguistics in a wider sense of the term), insofar as they deal with concrete languages in concrete times, trying to answer very specific (or *particular*) questions about their past, a fourth strand of research makes use of the new cross-linguistic databases along with results drawn from the phylogenetic approaches to investigate general aspects of language change, including questions like the rate of linguistic change and its correlates (Calude and Pagel 2011, Greenhill et al. 2017), the question to which degree environmental factors might have an impact on language evolution (Everett et al. 2015), or how language structures converge independent of contact or inheritance (Blasi et al. 2016).

Why is the aspect of dating, i.e., the inference of absolute phylogenies, so important for the new methods in historical linguistics?

Benefits of computational historical linguistics

Apart from the obvious benefit that the new quantitative methods have drastically revived the interest of scholars in historical linguistics, which also resulted in an increased amount of funding and a new generation of young scholars who are highly collaborative in their research and well trained in computational methods, the quantitative turn has also led to a considerable amount of rethinking in the field of historical linguistics, which offers new perspectives on the subject which have been ignored so far. First, we can see that the new methods shift the focus from *internal* to *external language* history, while at the same time turning away from the traditional focus on Indo-European alone.[3] We can also see that the new methods lead to the raise of new questions, specifically addressing *general* questions of language history.

This is also reflected in new research approaches, which are more explicitly *data-centered* nowadays and often based on statistical or stochastic modeling. While research in historical linguistics has always been data-centered, the new methods have shown that the classical approaches to deal with data – namely the individual collection of extensive personal notes from the literature, and the publication of new insights from these personal collections in form of extensive prose – are reaching their limits in times where the amount of data is constantly increasing. Although the attempts to automate the classical methods have so far not yet led to a situation where computers could beat the experts,[4] we

[2] See List (2019a) for an overview on these approaches.

[3] Compare classical handbooks such as the *Einführung in die vergleichende Sprachwissenschaft* by Szemerényi (1970), where the term *comparative linguistics* (which should be a general discipline) is seen as a synonym for *Indo-European linguistics*.

[4] This is also not to be expected shortly, given that the only areas in which machines outperform humans so far are restricted fields, such as chess, or the go-game (Silver et al. 2016), and not in problems that need to be solved in open worlds.

have won many important and new insights into the methods and the practice of historical language comparison, specifically also because the new methods challenged classical (traditional) linguists to revise the methods they use and to increase the degree of explicitness by which they apply them.

That languages interact with different factors is evident. What are the aspects that make it so difficult to study language change with help of computational frameworks?

Problems and criticisms

Not all linguists have enthusiastically welcomed the new methods. While the various critics range from justified criticism, via exaggerations, up to complete ignorance for the initial goals of the computational approaches, and at times rather reflect the insulted ego of those who consider themselves as indisputable experts, the new field faces a couple of serious problems that are worth being criticized and rigorously analyzed. Among the most important of these are (1) problems with the data that is used in quantitative analyses, (2) problems of applicability of the computational approaches, and (3) problems of transparency and (4) comparability with respect to the results and methods which scholars report, and (5) problems of the general accuracy of the computational methods in comparison with experts.

The data problems related to the way in which data are compiled and curated, and what judgments they are based upon. The general problem here is that most of the phylogenetic approaches still make use of human-annotated data, trusting the expertise of only a small amount of experts to be enough to annotated data for at times more than 100 different languages. The danger of this procedure (which is to some degree difficult to avoid) are potential problems of inter-annotator-agreement, which may themselves, of course, impact the results (Geisler and List 2010). The problem of applicability and transparency is reflected in large amounts of software solutions and datasets that are only discussed in the literature, but have not been openly shared (List et al. 2017). As a result, there are quite a few methods out there that could provide valid solutions, but which have only been tested on one dataset and never officially been published, which comes close to a crisis of irreproducibility as it has been noted in many branches of science since the beginning of this millennium (Nature 2013).[5]

The problem of comparability results from missing standards in our field, which make it difficult to compare results across datasets, since it is often very tedious to lift the data used by different scholars to a level where they could be easily compared. The problem of accuracy, finally, is probably the hardest problem to address, since the problems of historical linguistics are often quite hard to solve automatically, specifically also because – as a rule – data is sparse, while most computational methods have been built based on the assumption that data to test and train algorithms would be abundantly available.

What solutions can you think of to overcome the problems of transparency and comparability, which were mentioned above?

2 Towards a qualitative turn in diversity linguistics

Reconciling classical and computational research

The use of computer applications in historical linguistics is steadily increasing. With more and more data available, the classical methods reach their practical limits. At the same time, computer applications are not capable of replacing experts' experience and intuition, especially when data are sparse.

[5] Luckily, this picture is slowly changing, thanks to extensive efforts to propagate free data and free code. A our department, for example, we have now decided to refuse to review papers where we are not given code and data, if they are needed for replication, following the idea of referee's rights as expressed by the editorial board of the journal Nature in 2018.

If computers cannot replace experts and experts do not have enough time to analyse the massive amounts of data, a new framework is needed, neither completely computer-driven, nor ignorant of the assistance computers afford. Such computer-assisted frameworks are well-established in biology and translation. Current machine translation systems, for example, are efficient and consistent, but they are by no means accurate, and no one would use them in place of 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, a new paradigm of computer-assisted translation has emerged (Barrachina et al. 2008: 3).

Do you have experience with computer-assisted translation? If not, what role do computers and computer tools play for your research?

Computer-assisted language comparison

Following the idea of computer-assisted frameworks in translation and biology, a framework for computer-assisted language comparison (CALC) is 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 with human experts systematically correcting the results, they can drastically increase the efficiency of the classical comparative method and make up for the insufficiencies of of current computational solutions. At the same time, bringing experts closer to computational and formal approaches will also help to increase the consistency or classical research, forcing experts to annotated their specific findings and corrections in due detail, without resorting to texts in prose and ad-hoc explanations.

Classical linguists working on etymological research often emphasize the importance of looking into all details of language history, invoking the slogan “chaque mot a son histoire”, which is, according to Campbell (1999: 189) traditionally attributed to Jules Gilliéron (1854-1926). Even if this was completely true, how can we still defend the recent attempts of computer-assisted and computer-based strategies in historical linguistics to work on a more formal and more quantitative handling of linguistic data?

Data, Software, and Interfaces

In the framework of computer-assisted language comparison, data are constantly passed back and forth between computational and classical linguists. Three different aspects are essential for this workflow: Specific *software* allows for the application of transparent methods which increase the accuracy and the application range of current methods in historical linguistics and linguistic typology. Interactive *interfaces* serve as a bridge between human and machine, allowing experts to correct errors and to inspect the automatically produced results in detail. To guarantee that software and interfaces can interact directly, *data* need to be available in human- and machine-readable form.

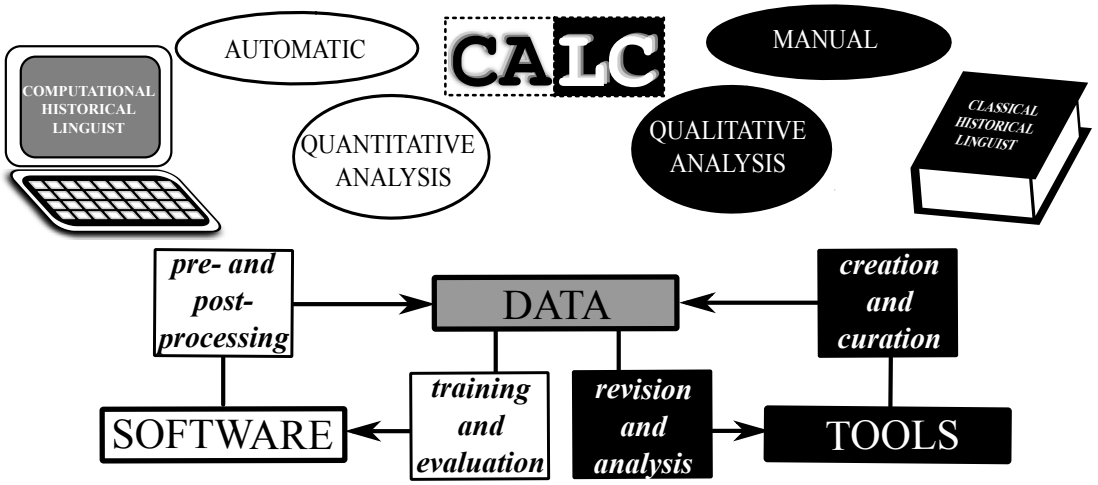

Fig. 1: Interplay of data, software, and interfaces in computer-assisted language comparison.

How exactly should one imagine data that are human- and machine-readable at the same time?

CALC project at the MPI-SHH in Jena

In the ERC-funded research project CALC (Computer-Assisted Language Comparison, List 2016), we try to establish a computer-assisted framework for historical linguistics. We pursue an interdisciplinary approach that adapts methods from computer science and bioinformatics for the use in historical linguistics. While purely computational approaches are common today, the project focuses on the communication between classical and computational linguists, developing interfaces that allow historical linguists to produce their data in machine readable formats while at the same time presenting the results of computational analyses in a transparent and human-readable way.

As a litmus test which proves the suitability of the new framework, the project attempts to create an etymological database of Sino-Tibetan languages (see Sagart et al. 2019 for initial attempts and results). The abundance of language contact and the peculiarity of complex processes of language change in which sporadic patterns of morphological change mask regular patterns of sound change make the Sino-Tibetan language family an ideal test case for a new overarching framework that combines the best of two worlds: the experience of experts and the consistency of computational models.

What may be the reason for choosing an interdisciplinary approach, and what are the most likely disciplines from which the project could take inspiration?

References

Atkinson, Q. D. and 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 and C. Renfrew. Cambridge, Oxford, and Oakville: McDonald Institute for Archaeological Research, 91–109.

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

Bergsland, K. and H. Vogt (1962). "On the validity of glottochronology." *Current Anthropology* 3.2, 115–153. JSTOR: 2739527.

Blasi, D. E., S. Wichmann, H. Hammarström, P. Stadler, and M. H. Christiansen (2016). "Sound–meaning association biases evidenced across thousands of languages." *Proceedings of the National Academy of Science of the United States of America* 113.39, 10818–10823.

Bouchard-Côté, A., D. Hall, T. L. Griffiths, and D. Klein (2013). "Automated reconstruction of ancient languages using probabilistic models of sound change." *Proceedings of the National Academy of Sciences of the United States of America* 110.11, 4224–4229.

Bowern, C., P. Epps, R. Gray, J. Hill, K. Hunley, P. McConvell, and J. Zentz (2011). "Does Lateral Transmission Obscure Inheritance in Hunter-Gatherer Languages?" *PLoS ONE* 6.9, e25195.

Calude, A. S. and M. Pagel (2011). "How do we use language? Shared patterns in the frequency of word use across 17 world languages." *Philosophical Transactions of the Royal Society B* 366, 1101–1107.

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

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

Donohue, M., R. Hetherington, J. McElvenny, and V. Dawson (2013). *World phonotactics database*. Canberra: Department of Linguistics. The Australian National University.
Dryer, M. S. and M. Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Everett, C., D. E. Blasi, and S. G. Roberts (2015). "Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots." *Proceedings of the National Academy of Sciences of the United States of America* 112.5, 1322–1327.
Geisler, H. and 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.
Gray, R. D. and Q. D. Atkinson (2003). "Language-tree divergence times support the Anatolian theory of Indo-European origin." *Nature* 426.6965, 435–439.
Gray, R. D., A. J. Drummond, and S. J. Greenhill (2009). "Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement." *Science* 323.5913, 479–483.
Greenhill, S. J., C. H. Wu, X. Hua, M. Dunn, S. C. Levinson, and R. D. Gray (2017). "Evolutionary dynamics of language systems." *Proceedings of the National Academy of Sciences of the United States of America* 114.42, E8822–E8829.
Grollemund, R., S. Branford, K. Bostoen, A. Meade, C. Venditti, and M. Pagel (2015). "Bantu expansion shows that habitat alters the route and pace of human dispersals." *Proceedings of the National Academy of Sciences of the United States of America* 112.43, 13296–13301.
Haspelmath, M. and U. Tadmor, eds. (2009). Berlin and New York: de Gruyter.
Hauer, B. and G. Kondrak (2011). "Clustering semantically equivalent words into cognate sets in multilingual lists." In: *Proceedings of the 5th International Joint Conference on Natural Language Processing*. (Chiang Mai, Thailand, 11/08–11/13/2011). AFNLP, 865–873.
Holman, E. W. et al. (2011). "Automated dating of the world's language families based on lexical similarity." *Current Anthropology* 52.6, 841–875. JSTOR: 10.1086/662127.
Hrozný, B. (1915). "Die Lösung des hethitischen Problems [The solution of the Hittite problem]." *Mitteilungen der Deutschen Orient-Gesellschaft* 56, 17–50.
Jäger, G. (2019). "Computational historical linguistics." *Theoretical Linguistics* 45.3-4, 151–182.
Kirby, K. R. et al. (2016). "D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity." *PLOS ONE* 11.7, 1–14.
Kolipakam, V., F. M. Jordan, M. Dunn, S. J. Greenhill, R. Bouckaert, R. D. Gray, and A. Verkerk (2018). "A Bayesian phylogenetic study of the Dravidian language family." *Royal Society Open Science* 5.171504, 1–17.
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.
Lee, S. and T. Hasegawa (2011). "Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages." *Proc. Biol. Sci.* 278.1725, 3662–3669.
List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.
— (2015). "Network perspectives on Chinese dialect history." *Bulletin of Chinese Linguistics* 8, 42–67.
— (2016). *Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics*. Jena: Max Planck Institute for the Science of Human History.
— (2019a). "Automated methods for the investigation of language contact situations, with a focus on lexical borrowing." *Language and Linguistics Compass* 13.e12355, 1–16.
— (2019b). "Automatic inference of sound correspondence patterns across multiple languages." *Computational Linguistics* 45.1, 137–161.
List, J.-M., S. J. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel (2018). "CLICS^2. An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats." *Linguistic Typology* 22.2, 277–306.
List, J.-M., S. J. Greenhill, and R. D. Gray (2017). "The potential of automatic word comparison for historical linguistics." *PLOS ONE* 12.1, 1–18.
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. Berlin, 599–605.
Maddieson, I., S. Flavier, E. Marsico, C. Coupé, and F. Pellegrino. (2013). "LAPSyD: Lyon-Albuquerque Phonological Systems Database." In: *Proceedings of Interspeech*. (Lyon, 08/25–08/29/2013).
Mennecier, P., J. Nerbonne, E. Heyer, and F. Manni (2016). "A Central Asian language survey." *Language Dynamics and Change* 6.1, 57–98.
Moran, S. and D. McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.
Nature, E. B. (2013). "Reducing our irreproducibility." *Nature* 496.4, 398.
— (2018). "Referees´rights." *Nature* 560, 409.
Nelson-Sathi, S., J.-M. List, H. Geisler, H. Fangerau, R. D. Gray, W. Martin, and T. Dagan (2011). "Networks uncover hidden lexical borrowing in Indo-European language evolution." *Proceedings of the Royal Society of London B: Biological Sciences* 278.1713, 1794–1803.
Prokić, J., M. Wieling, and J. Nerbonne (2009). "Multiple sequence alignments in linguistics." In: *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. "LaTeCH-SHELT&R 2009" (Athens, 03/30/2009), 18–25. acm: 1642052.
Sagart, L., G. Jacques, Y. Lai, R. Ryder, V. Thouzeau, S. J. Greenhill, and J.-M. List (2019). "Dated language phylogenies shed light on the ancestry of Sino-Tibetan." *Proceedings of the National Academy of Science of the United States of America* 116 (21), 10317–10322.
Silver, D. et al. (2016). "Mastering the game of Go with deep neural networks and tree search." *Nature* 529.7587, 484–489.
Swadesh, M. (1950). "Salish internal relationships." *International Journal of American Linguistics* 16.4, 157–167. JSTOR: 1262898.
Szemerényi, O. (1970). *Einführung in die vergleichende Sprachwissenschaft*. Darmstadt: Wissenschaftliche Buchgesellschaft.
Turchin, P., I. Peiros, and M. Gell-Mann (2010). "Analyzing genetic connections between languages by matching consonant classes." *Journal of Language Relationship* 3, 117–126.
Weiss, M. (2015). "The comparative method." In: *The Routledge handbook of historical linguistics*. Ed. by C. Bowern and N. Evans. New York: Routledge, 127–145.