Scientific Problem Solving

Johann-Mattis List (University of Passau)

1 Open Problems

Problems

When working every day on very detailed scientific problems, one always runs danger of loosing track of the broader challenges of one's field. That these challenges exist, and that we often still lack sufficient answers to certain problems becomes specifically clear when listening to the questions which laypeople or scientists from other fields ask with respect to one's area of expertise. In linguistics, for example, people are usually very surprised that the question of how language evolved the first time, the question regarding the origin of language, has been officially banned from the agenda of linguistics already in the 19th century, in the often-quoted *statuts* of the *Société de Linguistique de Paris*:

> La Société n'adment aucune communication concernant, soit l'origine du langage, soit la création d'une langue universelle. ("Statuts" 1871: III)

That there are in fact good reasons to avoid these questions becomes obvious when having a look at the large amount of speculative accounts on the origin of language, ranging from Herder's 1778 onomatopoetic speculation of early human beings running through the woods and imitating the sounds of the things surrounding them, or to recent mystic accounts, which have so far been ignored by a larger public:

> The Proto-Sapiens grammar was so simple that the sporadic references in previous paragraphs have essentially described it. The prime importance of sound symbolism for the people of nature should be noted again before we further detail that the vowel "E" was felt as indicating the 'yin" element, passivity, femininity etc., while "O" indicated the "yang" element, activeness, masculinity etc.; "A" was neutral or spiritual, indicating things conceived by the mind and emotions rather than with the physical senses. (Papakitsos and Kenanidis 2018: 8)

But at times, we may forget that there are valid problems in our field which we do not address, because we focus too much on the hard problems of the mainstream, or on tiny problems for which we know we might never find a sufficient answer. These problems may become evident when talking with laypeople, who may at times simply ask a question that would appear silly for a trained linguist. An example for such a question is the number of words that a language disposes of. While this sounds silly for linguists at the first sight, the question is in fact important for our science in multiple ways. It is important for the field of didactics, where it could help us to provide more efficient lessons on the most important words, it is important for historical linguistics, as it would allow us to measure how many of the words we can actually trace back in history, and it would be important for cognitive research, as it would allow us to assess the amount of information individuals can make use of when speaking.

In a paper on similarities between linguistic and biological evolution, we circumvented the question by giving a simple assessment on the words one needs in order to reach a level of proficiency according to different didactic studies (List 2016). But in the same year, Brysbaert et al. (2016) proposed a way to measure the amount of words that an English speaking person knows:

> Based on an analysis of the literature and a large scale crowdsourcing experiment, we estimate that an average 20-year-old native speaker of American English knows 42,000 lemmas and 4,200 non-transparent multiword expressions, derived from 11,100 word families. (ibid.: 1)

Starostin (1989) argues that every language has about 1000 roots which reflect its ancestry. Does this hold cross-linguistically and how much variation should we expect when comparing the languages of the world?

Hilbert and Hilpert Problems

At the and of the last year, inspired by a discussion I had with students who asked me about the biggest challenges for computational historical linguistics, I decided to sit down and make a short list of tasks that I consider challenging, but of which I think that they could still be solved some time in the nearer or further future. The idea to make such a list of questions is not new to mathematicians, who have their well-known Hilbert Problems, proposed by David Hilbert in 1900 (published in Hilbeert 1902). In linguistics, I first heard about them from Russell Gray, who himself was introduced to this by a talk of the linguist Martin Hilpert, who gave a talk on challenging questions for linguistics in 2014, called "Challenges for 21st century linguistics". Russell Gray since then has emphasized the importance to propose "Hilbert" questions for the fields of linguistic and cultural evolution, and has also presented his own big challenges in the past.

Due to my methodological background, the problems I identified and assembled are by no means big and in some sense also not necessarily extremely challenging (at least on first sight). Instead, the problems I decided for, when being asked, are problems I would like to see tackled, since I think they could help us to further advance our knowledge indirectly, by giving us the possibility to use the solutions of the problems to then answer deeper question on problems in multilingual computational linguistics. One further aspect of the problems that I selected is that these challenges can all be solved by algorithms or workflows. Even when being "small" in some sense, this does not mean, of course, that these problems are not challenging in the big sense. It also does not automatically mean that they can be solved in the near future. But given that the work in the field of computational and computer-assisted language comparison, progresses steadily, at times even at an impressive paste, I have some trust that these problems will indeed be solvable within the next 5-10 years.

What problems in your discipline do you consider unsolvable?

Ten Open Problems for Multilingual Computational Linguistics

When writing down my ten open problems for multilingual computational linguistics, I announced this in a blog post with the blog *The genealogical world of phylogenetic networks*, edited by David Morrison (`http://phylonetworks.blogspot.com/`), in January, with the plan of discussing each of the problems in detail in monthly blog posts throughout the year. At the end of 2019, ten problems had been discussed, and I later decided to elaborate further on them in order to write a small book. Until now, however, I have not found time to finish it or to make any significant progress in this regard.

The 10 problems, which are listed in Table **??** can be further classified into three different groups, which roughly correspond to three different categories important for research in general, namely *modeling*, *inference*, and *analysis*. This trias, inspired by Dehmer et al. (2011: XVII), follows the general idea that scientific research in the historical disciplines usually starts from some kind of idea we have about our research object (the *model* stage), and based on which we then apply methods to infer the phenomena in our data (the *inference* stage). Having inferred enough examples for the phenomenon, we can then *analyze* it qualitatively or quantitatively (the *analysis* stage) and use this information to update our model.

The first group in my list of problems deals with questions of *inference*, including the *detection of morpheme boundaries* (# 1), the *induction of sound laws* (# 2), the *detection of borrowings* (# 3), and

phonological reconstruction (# 4). What all these problems have in common is that they deal with inference in the sense described above, in so far as they start from linguistic data in some specific form, and the task is to find specific patterns in the data, which have not been annotated in the data beforehand.

The second group of problems deals with questions of *modeling*, including the *simulation of lexical change*, i.e., the design of consistent models that describe how the lexemes of a language change over time, the *simulation of sound change*, i.e., the simulation of the sound-change process by which sounds in a language change in dependence of the context in which they occur, and *the statistical proof of language relatedness*. While the simulation problems are clear problems of *modeling*, given that a simulation requires a model to be then applied to some artificial or existing datasets, the statistical proof or language relationship is a specific case, since it requires a model of language relatedness in order to test this model against a random model in which languages are thought to be unrelated. While there are numerous attempts in the literature to come up with a convincing statistical model to prove genetic relationship (Baxter and Manaster Ramer 2000, Kassian et al. 2015, Kessler 2001, Mortarino 2009, Ringe 1992), none of the attempts which have been proposed so far deals with lexical comparisons in all their complexity. Either, scholars only compare initial consonants with each other (Kessler 2001, Ringe 1992), or they resort to sound classes (Baxter and Manaster Ramer 2000, Kassian et al. 2015), and even if scholars compute random models for whole alignments of potentially related words (List 2014a), they have the problem of not accounting for the factor of closeness due to borrowing.

The last group of problems all have *typology* in their title, and belong to the class of *analysis* problems, dealing with the analysis of *semantic change*, *semantic promiscuity*, and *sound change*. What is meant by *typology* in this context is a data-driven estimate of the overall cross-linguistic frequency of these phenomena. Since we lack consistent accounts on the general tendencies of these processes and phenomena when excluding areal and genetic factors, the task is simply to come up with a consistent estimate on each of them. While semantic change and sound change are probably self-explaining in this context, the question of semantic promiscuity deserves some more attention. What is essentially meant by this term is the degree to which certain words, due to their original meanings, are re-used or re-cycled in the human lexicon.

While the term *promiscuity* has been used before in other contexts in linguistics, the specific usage of promiscuity to denote what one could also call *semantic productivity* or *concept productivity* was first proposed in List et al. (2016b), where biological and linguistic processes were consistently compared with each other, and semantic promiscuity was identified as a phenomenon similar to *domain promiscuity* in protein evolution in biology, with an explicit analogy being identified between the processes of *word formation* in linguistics and *protein assembly* in biology (ibid.: 5). For further elaborations of the concept of *semantic promiscuity*, compare List (2018) and Schweikhard (2018). Nowadays, I have again changed the terminology and no longer use the term *semantic promiscuity*, but rather *lexical root productivity* (List 2023).

Number	Problem	Class
1	automatic morpheme segmentation	inference
2	automatic sound law induction	inference
3	automatic borrowing detection	inference
4	automatic phonological reconstruction	inference
5	simulating lexical change	modeling
6	simulating sound change	modeling
7	statistical proof of language relatedness	modeling
8	typology of semantic change	analysis
9	typology of semantic promiscuity	analysis
10	typology of sound change	analysis

Does it seem useful to change one's terminology so often, and what may the rapid change in terminology for the same phenomenon reflect?

2 Computer-Assisted Strategies for Problem Solving

The way in which we carry out multilingual computational linguistics in this course is to follow a computer-assisted paradigm in which we try to design targeted methods that aid humans to do some boring tasks in a very accurate and efficient manner or to help humans to detect patterns in larger datasets. This means that we often have to develop new methods from scratch. In order to address the open problems in our field, some basic strategies for problem solving are helpful and important.

The framework for computer-assisted problem solving which I try to pursue in my own research and which I try to propagate does not neglect the possibility of using machine-learning techniques to tackle specific problems, but it does also not necessarily require that they be used exclusively. We do not naively accept machine learning solutions, but start instead from a careful inspection of the problems we actually want to solve. In many cases, a complex solution involving neural networks or Bayesian inference techniques may actually not be needed, since there are smart heuristics, or even complete solutions that do not require any stochastic component. In the same way in which we would not use a machine learning method to tackle the problem of multiplication, it is futile to have an algorithm searching for sound correspondences without any underlying model of sequence comparison or alignments.

That does not mean that machine learning solutions should be excluded per se, and in fact, many of the algorithms for cognate detection, which scholars call *supervised* or based on *linguistic knowledge*, make use of classical techniques, like random works, in specific stages of their workflow. But the decision when to use a specific technique is usually always based on some explicit reasoning that takes the phenomenon to be investigated into account, as well as the existing qualitative solutions that were developed within the field itself, and actual solutions in computer science or similar disciplines, such as bioinformatics, which are consulted to provide inspiration for possible solutions.

The current strategy, which has been applied to propose automatic solutions for various aspects of historical linguistics (List 2014b, List 2019) starts from a detailed investigation (also in collaboration with experts on the topic) of the existing qualitative solutions to a given problem in historical linguistics. As a second step, we try to describe the task in a clear way, by naming explicitly the input data and the output data we expect from the automatic method. We then try to model the process, while at the same time being prepared to further modify the requirements regarding the input data. The solution for the problem is then sought by looking at neighboring disciplines and topics, specifically graph theory, sequence comparison techniques in computer science and bioinformatics, in order to come up with a solution to the problem.

Does your discipline tend to use computer-based or computer-assisted approaches to tackle the major problems?

3 Modeling, Representation, and Implementation

Modeling and Representation

In the sciences, scholars often talk about *modeling*. Scholars *model* sound change, they *model* language change, and they try to *model* lexical borrowing. It is not always clear what is meant with the term *modeling*, and it seems that scholars use it with varying ideas in mind. If we talk about *modeling*

in the context of quantitative and formal approaches to historical language comparison, I use the term *model* in the sense of what Bröker and Ramscar (2021) call an *implemented model*. While a general model can also exist of a prose explanation of the mechanisms underlying a phenomenon, an *implemented model* is a model which can be shown to work in some piece of software and applied to some data.

> To explain why the contributions of representations, algorithms, and computations will only rarely manifest themselves in fully independent ways [...], it is important to recognise that in practice, models in the brain and cognitive sciences are typically presented in one of two distinct ways: either as abstract model descriptions, or as implemented models. Abstract model descriptions typically comprise symbolic (i.e. verbal or algebraic) descriptions of the relationships between what are typically quite loosely defined quantities or entities. Accordingly, while abstract models can appear to be more or less "formal", they typically fail to fully specify representations (what exactly will be counted and in which format) and typically fail to fully specify the algorithms that will transform these representations into predictions [...]. It is in fact only when these latter steps are made, and an abstract model is actually implemented, that it can be considered formal in any meaningful sense. (Bröker and Ramscar 2021: 17/25)

Of crucial importance for implemented models is the way in which data are *represented*, since this determines how the implementation works. In the work I will present, for example, we may conveniently represent language data (words in the lexicon of a language, etc.) in the form of tables. These can be printed to paper, but they can also be typed into spreadsheets on the computer. The representation of data is thus the basis upon which we build our models and implement them in computer code.

> To recapitulate: Representational choices can significantly alter the performance of a model, the predictions it makes and thus the way it is interpreted. (ibid.: 20/25)

The distinction between models, implemented models, and representations, does not define the term "model" itself. Atkinson and Gray (2006: 94) write about models, that they are "lies that lead us to the truth". Is this a useful characterization of models?

Integrated Data Representations

When working with data, scholars often use very different representations of their data. They may have one file for their syntactic properties they collect, one word document, where they collect their favorite quotes, another spreadsheet where they started to collect sound changes, and some old FileMaker database, which they still use for convenience, to enlarge their personal etymological dictionary of their favorite language. When working with data, scholars also often commit certain common errors in data collection. The most common errors are to extract information from sources without storing a reference to the original sources, or to copy text from some resource into a cell in a spreadsheet and later modify this content manually without keeping the original raw data.

As of now, there are many good guidelines for working transparently with data out in the internet (Perkel 2022), and I recommend that all who feel a bit insecure about how to collect data properly to inform themselves about these resources and generally take much more time in planning or experimenting with different formats of data representation than starting to collect data and eventually destroying information without having intended to do so. I also recommend to think about *integrated data representation*, that is, to think about ways to work on different questions with the same data, and to extract certain important aspects of the annotation of a dataset rather than paste it into a separate

data sheet. As an example, scholars may store a dictionary of a given language written in orthography, and additionally they may type off the phoneme inventory of that language from another resource and collect these separately. It would be much better to work on a dictionary in phonetic transcription from which the same information could be derived (the inventory should be extractable from the dictionary). Examples for integrated data handling have been recently published by our group in the form of the Lexibank repository (List et al. 2022), where we compute several lexical and phonological features of various languages from the wordlists, which we have collected and standardized.

Why is it so important to keep the raw data when collecting data for one's studies?

4 Modeling, Inference, and Analysis

Modeling

The models that are used so far in computational historical linguistics are all rather simple. While this may at times be surprising for classical linguists, who have a very complex idea of change process and also very detailed knowledge of the complex range of what is possible in language change, reducing the complexity of models is a necessary step in all scientific research. Rather then trying to establish the most complex models before we start to infer something, we should investigate how far we can go with a simplifying model and where its specific limits lie.

Crucial aspects for the models in multilingual computational linguistics are the concept of *language*, *word* (or linguistic sign), *word form*, and *word meaning*. Higher dimensions relevant for questions of language use, such as the speaker-listener interaction, are usually disregarded in the initial stages of investigation. The most common model for a language ist to treat a given language as a bag of words (or a bag of linguistic signs). Depending on the perspective, one can invoke a set of grammatical rules by which these signs are combined to form sentences. The linguistic sign itself follows the basic idea of Saussure (*Cours de linguistique générale*) with the modification that the sign is not seen as a duplet of *form* and *meaning*, but a triplet of form, meaning, and the *language* to which the sign belongs (List 2014b).

The sign form is usually modeled as a *sequence of sounds*, which implies that we can *segment* each word into a certain number of sounds. The sequences are constructed or constrained by *phonotactic rules*. If needed, one can add an additional layer of segmentation, dependent on the research question (e.g., one could look at a word consisting of morphemes consisting of sound segments, or a word consisting of syllables consisting of sound segments). These *secondary sequence structures* are of a certain importance in modern approaches for sequence comparison (List 2014b, List et al. 2016b), but they are often also deliberately disregarded. While the sign form is best treated as a sequence of sounds, the sign meaning is usually handled as a *network of senses*.

While this model of language as a bag of words may seem very simply, it is effectively the model that was underlying most of the phylogenetic analyses that have been published so far. Additionally one should say, that even classical historical linguists tend to use this model in their analyses. When needed, throughout this course, we will discuss more complex models in due time.

To address the problem that we face a drastic lack of comparability with respect to the data that has been produced in multilingual computational linguistics, the Cross-Linguistic Data Initiative (`https://cldf.clld.org`, Forkel et al. 2018) has published a set of recommendations for unified data standards in diversity linguistics, which are now gaining more and more popularity among scholars. These recommendations build more or less directly on the above-mentioned language model, and the current plan is to expand these further, based on the need and the availability of more complex models. As a very important aspect of standardization, CLDF comes along with *reference catalogs*, which

are basically meta-datasets, that offer standards for the handling of languages (Glottolog, https://glottolog.org, Hammarström et al. 2018), concepts (Concepticon, https://concepticon.clld.org, List et al. 2016a), and sounds in transcription (CLTS, https://clts.clld.org, Anderson et al. 2018).

In addition to the modeling of the data, the modeling of the processes, which has been not mentioned here, is of great importance. What models can you think of that would explain, for example, the process of sound change, or the process of lexical change?

Inference

As mentioned before, the inference of dated language phylogenies is by far the most popular of the computational methods proposed so far in the field of computational historical linguistics. Discussing the details of these approaches would, unfortunately, go beyond the scope of this session, but good review literature that provides some basic insights is now readily available (Greenhill 2015). What seems important to mention in this context is that the bag-of-words model mentioned before can be seen as the standard model that is essentially used to search for a language phylogeny. When discussing the simulation of language change in a later session, we will discuss more complex ways to simulate language change, which in theory also allow to handle the interaction between speaker and listener.

Second in popularity are methods for automated sequence comparison, which are very popular in dialectology, where methods for phonetic alignment are used to compute aggregate distances between dialect varieties, based on pronunciation distances derived from pre-selected lists of words (Nerbonne et al. 2011). In addition, methods for phonetic alignments are also used for the task of automated cognate detection, which tries to infer which words in a multi-lingual wordlist go back to the same ancestor. Techniques for automated cognate detection are quite well-developed by now, and have been shown to work surprisingly well, with accuracy scores of up to 90% on shallower language families (List et al. 2017), while the accuracy usually drops to around 60%-70% when dealing with larger datasets (Jäger et al. 2017). Further aspects of inference include automated borrowing detection (Mennecier et al. 2016), the detection of sound correspondences and sound correspondence patterns (List 2019), and also the automated prediction of so far unobserved words (Bodt and List 2019), which is specifically useful to support fieldworkers working on small groups of related languages.

How can automated word prediction be useful for linguistic field work?

Analysis

As it was mentioned briefly before, the distinction between what counts as inference and what counts as analysis are not always easy to draw. Intuitively, analysis should involve g-linguistic questions in the sense discussed in the first session, but it is clear that there is no formal justification for it, and it seems to depend more on the workflow, whether a certain step (such as – for example – phylogenetic inference) is labeled as part of the inference or the analysis step. An example for such a borderline case is the *Database of Cross-Linguistic Colexifications* (CLICS, https://clics.clld.org, Rzymski et al. 2020), which offers cross-linguistic accounts on polysemies, which are displayed in form of a network analysis that provides information on the relative cross-linguistic closeness of more than 1500 different concepts, reflected in more than 1000 of the world's languages. The more classical analyses which are usually presented, however, try to test certain theories by analysing the data which has been inferred previously. In these cases, the large-scale cross-linguistic databases, which are increasingly

produced, play an important role, as they allow scholars to test their hypotheses on a global scale, allowing them, for example, to test hypotheses regarding the transmission of Creole languages (Blasi et al. 2017), the evolution of syntax (Widmer et al. 2017), or the impact of our diet on evolution of our speech sounds (Blasi et al. 2019).

What hypotheses can be derived from historical linguistics that could be tested with the help of cross-linguistic approaches?

References

Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018). "A Cross-Linguistic Database of Phonetic Transcription Systems." *Yearbook of the Poznań Linguistic Meeting* 4.1, 21–53.

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.

Baxter, W. H. and 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, and L. Trask. Cambridge: McDonald Institute for Archaeological Research, 167–188.

Blasi, D. E., S. M. Michaelis, and M. Haspelmath (2017). "Grammars are robustly transmitted even during the emergence of creole languages." *Nature Human Behaviour* 1, 723–729.

Blasi, D. E., S. Moran, S. R. Moisik, P. Widmer, D. Dediu, and B. Bickel (2019). "Human sound systems are shaped by post-Neolithic changes in bite configuration." *Science* 363.1192, 1–10.

Bodt, T. A. and J.-M. List (2019). "Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa languages." *Papers in Historical Phonology* 4.1, 22–44.

Brysbaert, M., M. Stevens, P. Mandera, and E. Keuleers (2016). "How many words do We know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age." *Frontiers in Psychology* 7, 1116.

Bröker, F. and M. Ramscar (2021). "Representing absence of evidence: Why algorithms and representations matter in models of language and cognition." *Language, Cognition and Neuroscience* 37.1, 1–24.

Dehmer, M., F. Emmert-Streib, A. Graber, and A. Salvador, eds. and introd. (2011). Weinheim: Wiley-Blackwell.

Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (2018). "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics." *Scientific Data* 5.180205, 1–10.

Greenhill, S. (2015). "Evolution and Language: Phylogenetic Analyses." In: *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*. Ed. by J. D. Wright. Second Edition. Oxford: Elsevier, 370 –377.

Hammarström, H., R. Forkel, and M. Haspelmath (2018). *Glottolog*. Version 3.3. URL: http://glottolog.org.

Herder, J. G. (1778). *Abhandlung über den Ursprung der Sprache, welche den von der königl. Academie der Wissenschaften für das Jahr 1770 gesetzten Preis erhalten hat. Welche den von der Königl. Academie der Wissenschaften für das Jahr 1770 gesetzten Preis erhalten hat*. Berlin: Christian Friedrich Voß. Google Books: QP4TAAAAQAAJ.

Hilbeert, D. (1902). "Mathematical problems." *Bulletin of the New York Mathematical Society* 8.1, 437–479.

Jäger, G., J.-M. List, and P. Sofroniev (2017). "Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Long Papers*. "EACL 2017". Valencia: Association for Computational Linguistics, 1204–1215.

Kassian, A., M. Zhivlov, and G. S. Starostin (2015). "Proto-Indo-European-Uralic comparison from the probabilistic point of view." *The Journal of Indo-European Studies* 43.3-4, 301–347.

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

List, J.-M. (2014a). "Investigating the impact of sample size on cognate detection." *Journal of Language Relationship* 11, 91–101.

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

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

— (2018). "Von Wortfamilien und promiskuitiven Wörtern [Of word families and promiscuous words]." *Von Wörtern und Bäumen* 2.10. URL: https://wub.hypotheses.org/464.

— (2019). "Automatic inference of sound correspondence patterns across multiple languages." *Computational Linguistics* 45.1, 137–161.

— (2023). *Inference of Partial Colexifications from Multilingual Wordlists*.

List, J.-M., M. Cysouw, and 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. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Luxembourg: European Language Resources Association (ELRA), 2393–2400.

List, J.-M., R. Forkel, S. J. Greenhill, C. Rzymski, J. Englisch, and R. D. Gray (2022). "Lexibank, A public repository of standardized wordlists with computed phonological and lexical features." *Scientific Data* 9.316, 1–31.

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 (2016b). "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.

Mennecier, P., J. Nerbonne, E. Heyer, and F. Manni (2016). "A Central Asian language survey." *Language Dynamics and Change* 6.1, 57–98.

Mortarino, C. (2009). "An improved statistical test for historical linguistics." *Statistical Methods and Applications* 18.2, 193–204.

Nerbonne, J., R. Colen, C. Gooskens, P. Kleiweg, and T. Leinonen (2011). "Gabmap – A web application for dialectology." *Dialectologia* Special Issue II, 65–89.

Papakitsos, E. C. and I. K. Kenanidis (2018). "Going to the root: Paving the way to reconstruct the language of homo-sapiens." *International Linguistics Research* 1.2, 1–16.

Perkel, J. M. (2022). "Six tips for better spreadsheets." *Nature* 608, 229–230.

Ringe, D. A. (1992). "On calculating the factor of chance in language comparison." *Transactions of the American Philosophical Society*. New Series 82.1, 1–110. JSTOR: 1006563.

Rzymski, C. et al. (2020). "The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies." *Scientific Data* 7.13, 1–12. URL: https://clics.clld.org.

Saussure, F. de. *Cours de linguistique générale*. Ed. by C. Bally. Lausanne: Payot, 1916; German translation: — . *Grundfragen der allgemeinen Sprachwissenschaft*. Trans. from the French by H. Lommel. 2nd ed. Berlin: Walter de Gruyter & Co., 1967.

Schweikhard, N. E. (2018). "Semantic promiscuity as a factor of productivity in word formation." *Computer-Assisted Language Comparison in Practice* 1.11.

Starostin, S. A. "Sravnitel'no-istoričeskoe jazykoznanie i leksikostatistika [Comparative-historical linguistics and lexicostatistics]." In: *Lingvističeskaja rekonstrukcija i drevnejšaja istorija Vostoka* [Linguistic reconstruction and the oldest history of the East]. Vol. 1: *Materialy k diskussijam na konferencii* [Materials for the discussion on the conference]. Ed. by S. V. Kullanda, J. D. Longinov, A. J. Militarev, E. J. Nosenko, and V. A. Shnirelʹman. Moscow: Institut Vostokovedenija, 1989, 3–39; English translation: — . "Comparative-historical linguistics and lexicostatistics." In: *Time depth in historical linguistics*. Trans. from the Russian by I. Peiros. Vol. 1. Papers in the prehistory of languages. Cambridge: McDonald Institute for Archaeological Research, 2000, 223–265.

"Statuts" (1871). "Statuts. Approuvés par décision ministérielle du 8 Mars 1866." *Bulletin de la Société de Linguistique de Paris* 1, III–IV.

Widmer, M., S. Auderset, J. Nichols, P. Widmer, and B. Bickel (2017). "NP recursion over time: Evidence from Indo-European." *Language* 93.4, 799–826.