Problems in Assessing the Probability of Language Relatedness

Johann-Mattis List

Whether or more languages are genetically related and form a language family whose members all developed from a common source is one of the most frequently debated questions in the field of historical linguistics. In order to avoid being stuck in qualitative arguments about details, scholars have repeatedly tried to establish formal approaches that would allow to assess language relatedness in quantitative or probabilistic frameworks. Up to now, however, none of the numerous approaches which have been made so far has been able to convince a critical number of scholars. In this study, I will discuss common problems underlying all current approaches. By emphasizing that they all fail to provide sufficient estimates with respect to the *validity* of the test procedure, I will try outline common guidelines and tests that can help in the future development of quantitative approaches to the proof of language relatedness.

1 Introduction

Unlike in biology, where it is generally assumed that life on earth has only developed *one time* in the history of the planet, with all living creatures ultimately going back to a common ancestor, linguists are far more hesitant in assuming that language developed only once, and that all languages spoken on earth go back to the same source. There are different reasons for this hesitance. First, when assuming that first humans with similar capacities as we have them nowadays already evolved more than 100~000 (if not even more) years ago, we face a gap of more than 80~000 years which we cannot fill, since our acknowledged methods of historical language comparison do not seem to allow us to go farther back in time than about 20~000 if not less years. Second, in human history, we have witnessed that humans are capable of creating new languages *de novo*, as it is assumed to have happened with quite a few *signed languages*, although many share some common sources (Power, Grimm, and List 2020). Third, there is a tradition in the field of historical linguistics to deliberately *ignrore* the question of whether language evolved only once or multiple times, as it is presented most famously in the *statuts* of the Sociéte de Linguistique de Paris from 1866 ("Statuts: Approuvés Par Décision Ministérielle Du 8 Mars 1866" 1871):

> La Société n'admet aucune communication concernant, soit l'origine du langage, soit la création d'une langue universelle. [The society does not accept any kind of communication, neither on the origin of language, nor on the creation of a universal language]

Since linguists do not take a monophyletic origin of language for granted, they can not – unlike their colleagues from biology – start and compare all languages freely. Before they do so, they must sit down and search for sufficient proof that the languages they want to investigate comparatively share indeed a common ancestor. This task, commonly known as the *proof of genetic relationship* is so important for the field of historical linguistics, that numerous studies have been devoted to it.

While proving language relationship may be surprisingly simple at times, especially when witnesses in the form of written material are available, or striking similarities can be found in complex morphological paradigms, it may also be close to impossible, as reflected in at times century-long debates about the common or non-common origin about some highly contested language families, such as, for example, Altaic (Starostin 1991; Ceolin 2019).

The biggest problem for the proof of genetic relationship is that scholars themselves seem to have quite different opinions on those shared traits among languages that would count as evidence for relatedness. This discussion is often reflected in a supposed debate about *grammar* as proof on the one hand and *lexicon* as proof, on the other hand, although this reduction is in fact highly misleading, since *grammar* also refers to the *form* of the linguistic sign, and mostly reflects the idea that a solid proof of genetic relatedness should show similarities in inflectional morphology, or so-called *individual-identifying evidence* (Nichols 1996; see also Dybo and Starostin 2008).

In order to circumvent this problem, scholars have repeatedly tried to produce statistical tests that would help them to prove language relationship in a more objective manner. The basic idea of most of these tests is to show that it is highly unlikely that a certain similarity pattern, as it can be observed between two or more languages, has evolved by chance. Starting with the work by Ringe (1992), statistical tests of this kind have been repeatedly discussed, proposed, modified, and tested by various scholars in the past decades (Baxter and Manaster Ramer 1996, 2000; Mortarino 2004; Turchin, Peiros, and Gell-Mann 2010; Kassian, Zhivlov, and Starostin 2015; Ceolin 2019).

So far, however, no test has really seemed to convince linguists of being reliable enough to be either applied to a larger sample of language, or to be used as an argument to settle certain long-standing debates. On the contrary, it seems that despite all these tests we are even farther away from accepting any claims on more long-ranging relationships between languages than we might be without the tests, and standardized, large-scale approaches to language comparison have not been able to convincingly identify any deeper historical connections among the world's language families that could not also be attributed to former contact relations (Jäger 2015).

2 Initial Thoughts on Tests and Proofs

2.1 Absolute Proofs and Relative Tests

Classical historical linguistic scholarship treats genetic relationship in a manner similar to mathematicians who use proofs to mark a problem as solved. If the proof for genetic relationship has been identified (which may require a lot of genius), the problem is settled, and the reconstruction can begin.

The idea of designing a *test*, however, is fundamentally different in this regard, since tests only offer approximations to problems, and they are always accompanied by rates of false positives and false negatives. Designing a test to prove something is therefore an enterprise which is problematic already, since the idea of testing and the idea of finding a proof are fundamentally different.

This conflict is also one of the reasons for the numerous discussions among those who favor proofs and those who favor tests when it comes to dealing with genetic relationship. It explains why linguists who view genetic relationship as something which needs to be proven, react so harshly on statistical approaches that claim to *prove* something, and often refuse to take any quantitative arguments for granted.

Interestingly, it is not easy to find arguments neither against proofs nor against tests. It seems that both approaches deserve attention, and both have their pros and cons. Proofs are so demanding – since a proof needs to convince a multitude of scholars before it is widely accepted – that it is unlikely that we will ever reach a solid stage in some language families that scholars have proposed. In these cases, tests could be useful, specifically since they emphasize that our knowledge regarding language relationship is a question of degrees, even if we assume that genetic reflects a fact from an ontological perspective.

2.2 Test Theory and Construct Validity

Similar to linguistics, psychology and sociology often deal with phenomena which cannot be directly observed. Although we assume that attributes of people, such as "intelligence", exist in some sense, we cannot directly observe and measure them. While psychologists have a problem with meausuring intelligence, linguists have a problem with investigating proto-languages. Although few linguists would doubt that there was a common ancestral language from which Latin, Greek, Germanic, and other language groups evolved, we cannot provide any direct evidence in the favor of this assumption. Similar to historiographers, we cannot verify what happened in the past, the *res gestae* remain lost as long time travel is impossible (Schmitter 1982, 55f), and we have no way to verify *any* of the theories that have been proposed with respect to the structure of the Indo-European proto-language.

In order to cope with the problem of having to deal with research objects which are not directly accessible to investigation, psychologists have long since developed the concept of *constructs*. A construct refers to something that is not directly observable, but that may leave traces in tests (Cronbach and Meehl 1955, 178). While psychologists tend to believe that the concept of *intelligence* is real, they agree that they cannot really investigate it, and what they investigate instead is their *construct of intelligence*. Constructs can be defined in different ways, and some constructs of intelligence will perform better than others, when putting them to the test. We can also assume that constructs will evolve over time and become more sophisticated, with more scientists collaborating to enhance them.

While the notion of constructs is not common in the field of historical linguistics, the notion seems useful, specifically when looking at the numerous debates in which linguists have been discussing the “nature of the proto-language”, with one camp maintaining the view that our reconstructions are basically “realistic” and one camp emphasizing that reconstructions are mere abstractions. It seems the whole debate was misleading all the time, since the question is a question of degree, similar to the constructs in psychology. As long as we trust that our methods truly identify a proto-language, we will assume that this proto-language is an *ontological fact*, but how we approach this proto-language, what we can learn about it, how we can reconstruct it, depends on the *epistemological reality* (Kormišin 1988, 92), and here, the question of abstract versus realistic reconstructions is more a question of degree between two extremes. Our linguistic reconstruction thus addresses our *construct of the proto-language*, and this construct is the “fiction or story put forward by a theorist to make sense of a phenomenon” (Statt [1981] 1998, 67), it is not the real object, not the real language which was once spoken (see also List 2014, 51–54).

In order to assess, how well a given test measures a given phenomenon, scholars in psychology and sociology make use of the notion of *construct validity*, apart from *objectivity* and *reliability* one of the three control criteria for scientific experiments. Construct validity describes “[...] how well each item of a [...] test measures or predicts what it’s supposed to measure or predict” (Statt [1981] 1998, 30). In a more theoretical notion, validity refers to the “strength” of theories and hypotheses (Liebert and Langenbach Liebert [1973] 1995, 100–119). Although validity is sometimes regarded as an absolute value, it is important to note that the concept is gradual in its nature (otherwise one would have to conclude that a fast watch could not measure time at all). This is in concordance with the common truth theories in philosophy, which likewise do not claim absolute truth as a practicable goal of scientific research (Zoglauer 2002, 27–32). If one subscribes to this gradual notion of validity, one can conclude that validity designates a theory’s “closeness” to reality, i.e. the degree to which a theory reflects a certain reality. Validity in this sense can be seen as the relation between two sets, one comprising a theory, the other a “reality”.

It is not easy to assess the validity of a given test, and there is a lot of debate in the social sciences on this topic. Similar to linguistic debates about the nature of the proto-language, the question to which degree a certain construct such as intelligence can be measured after all, often also has a philosophical dimension, where formal arguments no longer apply. Validity itself depends on *reliability*, referring to the degree to which “a particular observation has yielded a replicable score” (Liebert and Langenbach Liebert [1973] 1995), and objectivity “kennzeichnet die Unabhängigkeit seines Ergebnisses von der Person, die den Test durchführt” (“refers to the independence of the result from the person conducting the test”; Maderthaner 2008, 89).

Given that scholars often treat the proof of genetic relatedness as a specific *test*, the general control criteria developed in the test theory of the social sciences should also apply to them. As a result, we can use the notion of constructs and construct validity and try to assess the quality of proposed tests. Furthermore, when working on tests for genetic relationship in the future, it will be helpful to keep these basic criterias in mind.

3 How to Design Tests for Relatedness?

If we get back to the three control criteria for tests in the social sciences, objectivity, reliability, and validity, it is useful to start by identifying what these criteria could mean for tests of genetic relatedness in concrete.

3.1 Objectivity

With respect to objectivity, referring to the independence of a test from the person applying the test, we can see that we already enter thin ice in the context of historical linguistics, where we look back at a long history of celebrated individual scholarship that uses intuition acquired in years of painstaking labor to find the truth about a word's history, rather than relying on formal procedures that could be applied by any person following the instructions. However, it should be clear that science cannot rely on the intuition of experts alone, even if this position is still occasionally favored in the field of historical linguistics. As a result, when trying to make sure that our tests for genetic relatedness are objective, we need to make sure that we test to which degree the person conducting the test could have an influence on the results.

If we want to guarantee objectivity, there are two basic approaches that we can follow. First, we can automate our test procedure to such a degree that we no longer need any person conducting the test. Second, we can ask several colleagues to conduct the same test and see to which degree they converge in their assessments. While the advantage of fully automated frameworks is obvious, it is much less simple than one might think to really guarantee that a test is fully objective in this regard. Although methods for automated cognate detection have been constantly improving lately (List, Greenhill, and Gray 2017, @List2019a), and fully automated workflows for cognate detection and phylogenetic reconstruction have been proposed (Jäger 2015, Wu2020) and tested with quite some success (Rama et al. 2018), one should not forget that the automated comparison is usually only *one* part of a test for genetic relatedness. The first and often even most important part consists in the selection of the comparanda, the comparative concepts (Haspelmath 2010) by which the languages should be compared, and it has been shown that this part is among the least reliable ones of the whole test procedure (Geisler and List 2010).

In order to attest truly, how objectively a given test for genetic relatedness can be applied, it therefore seems impossible to do so without the one or the other test for inter-annotator agreement. Especially when it comes to the selection of the comparanda, it is indispensible to measure to which degree individual choices could impact the results of tests,and although it is not clear to which degree this might be the case, we should not ignore this aspect of the procedure.

3.2 Reliability

While objectivity depends on the person who conducts the test, reliability refers to the ability of a test to yield consistent results when applied in different situations. Since there are many different *situations* which may be different when re-applying a test, it is not easy to give a clearcut assessment of how to assess the reliability of a given test completely. We can, however, try to identify at least a minimal amount of aspects in which we would require a given test for genetic relatedness to yield similar results.

As a first criterion for reliability of tests for genetic relatedness, one could require that a test which shows that languages like German and English are related should equally be able to show that Mandarin Chinese and Cantonese are related. One may think that this critieron is trivial, but one should not forget that in the light of claims that *only* individual-identifying evidence proves relatedness (Nichols 1996), this criterion for a good, reliable test, does not immediately hold. While we find – to some degree – paradigmatic morphology recurring in English and German, we will have a hard time to do so when comparing Cantonese and Mandarin Chinese. One could of course also explicitly emphasize that a certain test for relatedness only works on inflecting languages. In any case, however, it is important that a test explicitly states to which languages it can be applied, and that a test *is* applicable to a coherent set of languages, and it is also clear that scholars may tend to prefer those tests that work on all spoken languages.

As a second criterion for a sufficiently reliable test for genetic relatedness, one should set up a certain time depth until which a given test should work at least. Since it is very likely that there is a definite maximum time depth until which we can identify regular sound correspondences or other kinds of formal similarities among languages, it is clear that every test has a certain and definite time depth. As a result, a test should come along with an approximate estimate of this time depth. Otherwise, the test would be useless when applying it to some long-range-proposal. The problem of applying tests for shallow time depths to long-range data is that neither a positive nor a negative result will bring us any further. In the case of a positive result, scholars won't be convinced anyway, but even more problematically, scholars who are skeptical of the proposal could easily think that a negative results confirms their assumption.

It is difficult to test the reliability of a test for genetic relatedness, especially when the initial procedure of data preparation is tedious, but in any case, scholars who propose tests should make sure that they provide an extensive gold standard, which is also dividied into different time depths, in order to make sure that they rigorously test how well their test performs across different language families at different times.

3.3 Validity

Validity is the hardest criterion to test when dealing with tests for genetic relatedness. At this point, I cannot provide any clear recommendations on how to make sure that a test is valid. But if we get back to the often repeated idea that validity tells us if a test measures what it is supposed to measure, it is very important that we make sure that our tests really assess genetic relationship among langues and nothing else. In this context, it is important that we remind ourselves of the four basic kinds of similarity which linguists are confronted with when comparing languages as repeated in the graphic below (from List 2014, 56).

What we can see here is that genetic similarities are only one out of four general types of similarities, namely "coincidental similarities" (which we can always encounter when comparing languages), "natural similarities" (which are due to common typological tendencies), and "contact-induced similarities" (due to language contact). Thus, if we design a test for genetic relatedness,

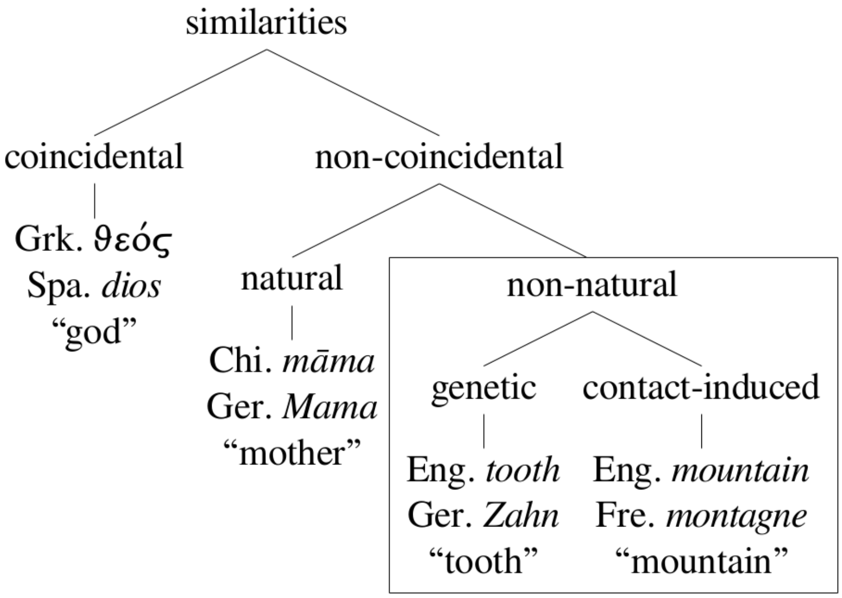

Figure 1: Four aspects of similarity in linguistics

we *need* to make sure that we do not measure any other kind of relatedness. It is obvious that this is very difficult, and scholars have tried to circumvent the problem in many ways, be it by comparing traits that are difficult to borrow, or by comparing language pairs where language contact can be ideally excluded.

4 Reviewing Tests in the Literature

When summarizing what was discussed above, we can find the following minimal requirements for any test of genetic relatedness that should be taken seriously by historical linguistics:

- [1] the test should either be fully automated or otherwise it should come along with extensive tests on inter-annotator agreement
- [2] the test should be applied to a large gold standard of different languages from different time depths
- [3] the test should be able to cope with contact-induced and natural similarities (e.g., sound symbolism)

When reviewing the literature that has dealt with tests for relatedness in the past, we can see that most tests try to cope with some of the challenges mentioned here, but it is also clear that there is *no* single test that tries to account for all three criteria.

4.1 Correspondence-Based Approaches

A first class of tests which were proposed so far can be called "correspondence-based approaches", since they try to identify matches in wordlists based on sound correspondences rather than similar sounds. A first study to mention in this context is Ringe (1992), who used 200-item wordlists (based on the concepts by Swadesh 1955), and then compared languages in a pairwise fashion and then derived a correspondence table from the initial sounds in each wordlist. The statistics were criticized by Baxter and Manaster Ramer (1996) and later refined by Kessler (2001), with a recent example by Ceolin (2019). The procedure consists of stages:

1. Compile a wordlist for all languages by translating the items of the concept list into the target languages, restricting translations to only *one* word.
2. Make a matrix of *n* sounds occuring as initials in language A and *m* sounds occurring as initials in language B, and fill in in how many cases each sound pair cooccurs.
3. Evaluate the results statistically.

I deliberately ignore *how* the statistical evaluation is carried out in the different versions of these tests, since these are less important for the kind of critic that is provided in this study. All in all, when only looking at our check list for good tests that conform to our control criteria, we can already identify several problems underlying these studies so far. These relate to (1) the procedure by which the wordlists for the languages are compiled, (2) the amount of languges on which this approach was tested, (3) the lack of explicit attempts to control for borrowing or sound symbolism (or to estimate its impact), (4) the model of relatedness, and (5) the general *power* of the test.

Regarding (1), we face a problem of objectivity, since the compilation of wordlists is not unproblematic, specifically, when synonyms are not allowed and scholars have to choose among them in cases of doubt (Geisler and List 2010). As an earlier study by List (2018) has shown, the impact of word choice on phylogenies and other aspects should not be underestimated, as shown in Figure 2, where the variation resulting from randomized choices among synonyms is displayed in the form of a consensus network. When dealing with wordlists, it is important to test how much different annotators would differ in the choice of words from different sources, to allow us to estimate how much this could influence the results of the test.

Regarding (2), we see a lack of reliability since the methods were only tested on very few languages so far, although by now, enough data should be available. The lack of gold standard testing becomes even more problematic when considering studies with negative results, such as the one by Ceolin (2019). Since we have no clear idea what it means if the test fails, it is impossible to use the study as an argument *against* the Altaic hypothesis. What if the study also fails on Indo-European? This would mean nothing but that the method is not reliable. Negative tests only bear information if we know how often we can expect them.

Since no control for borrowing or sound symbolism has been undertaken, and not attempts were made to include them in the model, it is not clear to which degree the method might suffer from the impact of language contact or parallel evolution in lexical change. All scholars argued in their studies, that their results are unlikely to be obtained by chance, but it is very difficult to really asses what *chance* would mean, specifically since we also know that sound symbolism should not be underestimated (Blasi et al. 2016).

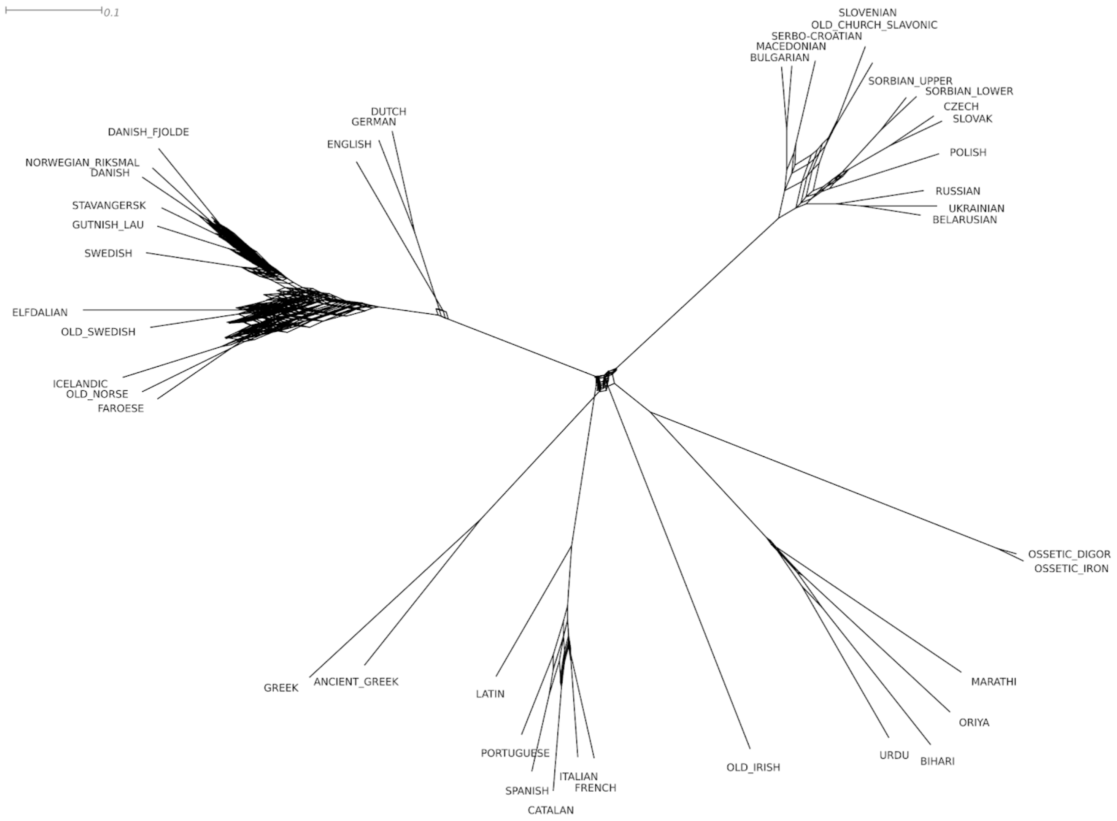

Figure 2: Consensus network of phylogenetic trees based on different synonym choices

Since the approach is based on semantically aligned wordlists, the tests take only potential cognate words with the same meaning into account, which reflects an extremely simplified model of relatedness (4). This yields, along with the fact that only the initial sounds are considered, and no true alignment of the words is carried out, a situation in which the test is deprived of much of its potential power (5), and it also explains why the tests perform so badly on not so distantly related languages.

4.2 Sound-Class-Based Approaches

Ultimately going back to Dolgopolsky (1964), sound classes have been employed by many scholars as a way to handle extreme phonetic and phonological variation and to allow for an exploratory analysis of language data in search for potential cognates. Starting with Baxter and Manaster Ramer (2000), Dolgopolsky's approach was further supplemented by statistical tests (usually based on permutation), and then also applied to more languages by Turchin, Peiros, and Gell-Mann (2010) or proto-languages (Kassian 2015).

The workflow consists of the following three steps:

1. Compile a wordlist (usually a version of Swadesh 1955) and extract consonant classes for the target languages.
2. Count the number of *matching* consonant classes for each language pair.
3. Assess the assumed significance of the matches by applying permutation tests.

As in the case of correspondence-based approaches, we can identify some obvious problems in theses approaches, of which some are specific and some are general. First, the wordlist compilation is as problematic as in the case of the correspondence-based approaches, and one may even argue that it is slightly more prone to errors, since the conversion *to* sound classes requires usually at least some degree of manual intervention. Second, although the method has been tested on *more* data, gold standard testing was still lacking in many studies, which is specifically problematic when dealing with proto-languages and testing only one pair which is presented to be significant (Kassian 2015). Third, while the scholars usually assume that the choice of concepts prevents problems resulting from borrowing or sound symbolism, the same problems apply to sound-class-based approaches as well, especially since we do not have global borrowing statistics (List 2019a). Fourth, as in the case of the correspondence-based approach, the approach deliberately ignores the process of semantic shift. Fifth, studies that compare how well sound-class-based approaches perform in identifying cognates have shown that they are rather reliable with respect to a low amount of false positives, while they also miss many cognates, and thus show a larger amount of false negatives (List, Greenhill, and Gray 2017).

5 Conclusion

I have provided a potentially rather skeptical overview on my personal view on tests for genetic relatedness. While this view is predominantly pessimistic with respect to research that was published, I hope that what I presented hear can have a positive impact on the development of future methods. What I wish for are:

- [A] methods that are based on open research paradigms, with open data, and open code, so that others can re-apply, newly apply, and build on them,
- [B] methods that are exhaustive, being tested on more data, not just on a couple of languages in which the scholars are interested, and
- [C] methods that are explicit on their limitations and provide exhaustive statistics on their success and failure on a gold standard.

When discussing language relatedness in the light of test theory and construct validity, there is one obvious problem, however, where the parallel to "testing" as we know it from medicine and psychology, does *not* hold. When looking back at the history of Indo-European linguistics and historical linguistics, we do not see that people *tested* that Sanskrit and Greek are related, but rather that they *detected* this relationship. Scholars happened to compare the right traits and to find evidence as striking as a video proof conducting a murderer. Often, scholars still think about relatedness in absolute dimensions, not in relative dimensions. I think, in the future, we need to advance in both directions: we need to strengthen our heuristics *and* we need to strengthen our tests. # References

Baxter, William H., and Alexis Manaster Ramer. 1996. "Review: On Calculating the Factor of Chance in Language Comparison. By Donald A. Ringe, Jr. Philadelphia: The American Philosophical Society, 1992. Pp. 110." *Diachronica* 8 (2): 371–84.

———. 2000. "Beyond Lumping and Splitting: Probabilistic Issues in Historical Linguistics." In *Time Depth in Historical Linguistics*, edited by Colin Renfrew, April McMahon, and Larry Trask, 167–88. Cambridge: McDonald Institute for Archaeological Research.

Blasi, Damián E., Søren Wichmann, Harald Hammarström, Peter Stadler, and Morten 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): 10818–23.

Ceolin, Andrea. 2019. "Significance Testing of the Altaic Family." *Diachronica* 36 (3): 299–336. https://doi.org/https://doi.org/10.1075/dia.17007.ceo.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52: 281–302.

Dolgopolsky, Aron B. 1964. "Gipoteza Drevnejšego Rodstva Jazykovych Semej Severnoj Evrazii S Verojatnostej Točky Zrenija." *Voprosy Jazykoznanija* 2: 53–63.

Dybo, Anna, and George S Starostin. 2008. "In Defense of the Comparative Method, or the End of the Vovin Controversy." In *Aspekty Komparativistiki*, edited by I. S. Smirnov, 3:119–258. Moscow: RGGU.

Geisler, Hans, and Johann-Mattis 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*, edited by Heinrich Hettrich. Wiesbaden: Reichert.

Haspelmath, Martin. 2010. "Comparative Concepts and Descriptive Categories." *Language* 86 (3): 663–87.

Jäger, Gerhard. 2015. "Support for Linguistic Macrofamilies from Weighted Alignment." *Proceedings of the National Academy of Sciences of the United States of America* 112 (41): 12752–7.

Kassian, Alexei. 2015. "Towards a Formal Genealogical Classification of the Lezgian Languages (North Caucasus): Testing Various Phylogenetic Methods on Lexical Data." *PLoS ONE* 10 (2): e0116950.

Kassian, Alexei, Mikhail Zhivlov, and George S. Starostin. 2015. "Proto-Indo-European-Uralic Comparison from the Probabilistic Point of View." *The Journal of Indo-European Studies* 43 (3-4): 301–47.

Kessler, Brett. 2001. *The Significance of Word Lists: Statistical Tests for Investigating Historical Connections Between Languages*. Stanford: CSLI Publications.

Kormišin, I. V. 1988. "Prajazyk. Bližnjaja I Dal'njaja Rekonstrukcija." In *Sravnitel'no-Istoričeskoe Izučenie Jazykov Raznych Semej*, edited by Ninel´Z. Gadžieva, 3:90–105. Moscow: Nauka.

Liebert, Robert M., and Lynn Langenbach Liebert. (1973) 1995. *Science and Behavior. An Introduction to Methods of Psychological Research*. Englewood Cliffs: Prentice Hall.

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

———. 2018. "Tossing Coins: Linguistic Phylogenies and Extensive Synonymy." *The Genealogical World of Phylogenetic Networks* 7 (2). http://phylonetworks.blogspot.de/2018/02/tossing-coins-linguistic-phylogenies.html.

———. 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* 1 (45): 137–61. https://doi.org/10.1162/coli_a_00344.

List, Johann-Mattis, Simon J. Greenhill, and Russell D. Gray. 2017. "The Potential of Automatic Word Comparison for Historical Linguistics." *PLOS ONE* 12 (1): 1–18.

Maderthaner, Rainer. 2008. *Psychologie*. Köln; Weimar; Wien: UTB.

Mortarino, Cinzia. 2004. "A Statistical Test Useful in Historical Linguistics." In *Proceedings of the XLII Scientific Meeting of the Italian Statistical Society*, 107–10.

Nichols, Johanna. 1996. "The Comparative Method as Heuristic." In *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, edited by Mark Durie, 39–71. New York: Oxford University Press.

Power, Justin M., Guido Grimm, and Johann-Mattis List. 2020. "Evolutionary Dynamics in the Dispersal of Sign Languages." *Royal Society Open Science* 7 (1): 1–30.

Rama, Taraka, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. "Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics?" In *Proceedings of the North American Chapter of the Association of Computational Linguistics*, 393–400.

Ringe, Donald A. 1992. "On Calculating the Factor of Chance in Language Comparison." *Transactions of the American Philosophical Society*, New series, 82 (1): 1–110. http://www.jstor.org/stable/1006563.

Schmitter, Peter. 1982. *Untersuchungen Zur Historiographie Der Linguistik: Struktur – Methodik – Theoretische Fundierung*. Tübinger Beiträge Zur Linguistik 181. Tübingen: Gunter Narr.

Starostin, Sergej Anatolévic. 1991. *Altajskaja Problema I Proischoždenije Japonskogo Jazyka [The Altaic Problem and the Origin of the Japanese Language]*. Moscow: Nauka.

Statt, David A, ed. (1981) 1998. *Consise Dictionary of Psychology*. 3rd ed. London; New York: Routledge.

"Statuts: Approuvés Par Décision Ministérielle Du 8 Mars 1866." 1871. *Bulletin de La Société de Linguistique de Paris* 1: III–IV.

Swadesh, Morris. 1955. "Towards Greater Accuracy in Lexicostatistic Dating." *International Journal of American Linguistics* 21 (2): 121–37. http://www.jstor.org/stable/1263939.

Turchin, Peter, Ilja Peiros, and Murray Gell-Mann. 2010. "Analyzing Genetic Connections Between Languages by Matching Consonant Classes." *Journal of Language Relationship* 3: 117–26.

Zoglauer, Thomas. 2002. *Einführung in Die Formale Logik Für Philosophen*. 2nd ed. Göttingen: UTB.