**Exploring and dating 8.000+ years of language relationships using statistical methods**

A quantitative approach to comparative linguistics

Vincent Beaufils

Independent researcher

Vienna, Austria

[vincent@beaufils.at](mailto:vincent@beaufils.at)

**Retracing the origin of our languages and building their family trees gives a fascinating glance in the past. Matching data and facts from the fields of genetics, archaeology and linguistics is the best way to stabilize hypotheses about human prehistory, yet many linguistic classifications and chronologies are still subject of debate. Quantitative and computerized approaches remain an underutilized option in comparative linguistics – probably because of the controversies associated with lexicostatistics and glottochronology in the past decades. Things have begun to change in the last 15 years, with new approaches for quantitative comparative linguistic yielding strong results.**

**In this paper, we present an approach using a very limited amount of roots as signals to identify and quantify language relatedness. The items selection relies on search for stability: semantically, against borrowing and against erosion. This search is based on statistical methods, which quantify not only relatedness signals items yield, but also their exposure to chance. The small amount of roots retained is the result of this statistical selection process. To compensate the low amount of appropriate roots, the system does not only estimates if pairs of lexical items in different languages are cognates or not, but also to which degree. The quantification of the level of erosion within the words delivers the signals, which lead to a phylogenetic classification and a chronology for the major language families of the world.**

**Chance interference plays a significant role in quantitative linguistic studies. For this reason, we apply chance calculation to every relatedness signal. It helps to identify which signals are helpful – it is not about *justifying* results – it is about *sharpening* them: managing the trade-off between “signal relevance” and “chance interference” is one of the key assets throughout the project.**

**After exploring the methodology and the data basis, we show that they lead to a reliable language classification across multiple language families and a strong hypothesis for dating proto-languages and placing the phylogenetic results in a chronological context. An attempt to explore more ancient relationships leads to statistical evidence for a part of the - still disputed – Eurasiatic and Nostratic macro families.**

# Introduction

The basis data of this paper has been compiled and made available by the author on a blog**1** since February 2013. It can be browsed online with queries of pairwise comparisons as well as in phylogenetic trees generated from the results. It was not originally designed for research but it has been growing overtime, with the results being challenged by various linguists or simply by native speakers from around the world. This dialog has led to improvements. Because the data delivers a good classification for most world languages, the idea is now to go into details and apply a more scientific approach, which is the subject of this paper.

The heart of the system is an algorithm, which calculates the distance between any two languages. A constant set of rules, which is applied equally to all comparisons, delivers a value between 0 and 100. For this, a set of 18 items from basic vocabulary has been selected. The algorithm processes a relatedness judgement out a set of rules for universal sound change. The probability that a positive judgement is due to chance is calculated for each pairwise comparison. The system produces an output for any pairwise comparison and a distance matrix containing all results.

It is a clear objective to use a transparent and stable set of rules to let the computer identify cognates and quantify their degree of erosion. The rules are applied the same way, regardless of the language family and with no exception. Algorithm-based judgments follow transparent, systematic and documented patterns and can thus be better addressed in statistical models than human based judgements.

Browsing the pairwise comparisons brings approximately 94% results which are in line with known classifications**2**. The real value of the data does not lie in the single comparisons, but in the complete result set which is the subject of this paper. Processing the data as a distance matrix and a set of distributions allows classifications and regression analyses, which lead to phylogenetic trees and evolution timelines.

# Choice of the lexical basis

The Swadesh list, a standard list of basic vocabulary, is considered as the first reference for comparative linguistics based on lexical items. Since its first release with 215 words in 1955**3** it has been a constant trend to reduce the number of words, concentrating on the most stable ones, with a final version of 100 words by Swadesh himself in 1972**4**. [Sergei Yakhontov](https://en.wikipedia.org/wiki/Sergei_Yakhontov) has established a 35 subset**5** and more recently, the ASJP project uses a 40 words list for a worldwide classification**6**. For this project, because it was the aim to quantify language relationships as far as possible in time, the choice has been made on following criteria:

* the words should stand for concepts which were supposedly relevant to most human populations as early as 6000 to 8000 years ago and more,
* they have to be stable against borrowing - known borrowings have to be excluded,
* the resistance to semantic shift is a criterion for the choice of the words. When a shift is identified, possible synonyms or derivational drifts have not been used, as the knowledge of these is not equally available in all language families. Applying changes only to some language families would bias the results.

*The selection process is not manual or subjective*

The above criteria are just a starting point for selecting words in basic vocabulary lists. The further process of selecting the word list for the study is not manual. It is, in fact, very similar to the idea behind supervised machine learning, where algorithms learn the response variable against a provided set of predictor variables**7** : applying several word lists to generate already known language classifications until a word combination generates the one that best matches known classifications. We try to reach the best error reduction (error between inferred classifications and existing ones) and to do so, reduce or increase the size of the word list. At the end of the process, the items with a similarity detection power, which is stronger than their exposure to chance, remain in the selection. Whenever an item biases the results, it means that it is not stable enough and/or its exposure to chance is stronger than the potential signals it yields toward relatedness detection. This process leads to a final set of roots, which is very limited. However, this limitation of the number of items is not arbitrary: it is the result of the error reduction process on a training set**8**: the Indo-European classification. As in machine learning, a validation set**8** is necessary to verify if the optimal word combination obtained also performs well on other language families. The results of the system in the classification of other know language families represents the validation set.

The set of 18 words presently used is the one, which performed best. Later attempts to add or remove words to sharpen the results proved to bring improvements, if any, only in some of the short-range classifications, with more chance interference in medium and long ranges. The fact that more words do not mean better classifications has become clearer and clearer throughout the selection process. Other researchers conclude that only very few words have a high degree of stability against borrowing, semantic drift and phonological erosion**9**.

*Words have different powers to retrace relatedness*

Each word has a distinct capacity to identify relatedness – especially with regard to time range. It does not mean that certain words have a general better performance than others: they simply reflect different aspects, which are more or less effective to identify relatedness in different periods: some words are more useful for short range classification. If we want to focus on short-range classifications, a longer word list performs better. In this study, our goal is to infer longer-range classifications of languages. This goal determines the choice of lexical items used in the classification.

*Long-range signals are typically more complex to trace*

The search for long-range relatedness of languages applied by famous authors delivers an abundant evidence for connections between languages beyond the universally recognized macro families**10 11 12 13**. The longer the comparison range, the more indirectly signals come: relatedness in synonyms or in semantically shifted meanings deliver a significant part of the evidence. However, in these cases, the exposure to chance is bigger and more difficult to measure, opening the doors for critical judgements and discussions, which are difficult to answer. It is very easy to refute the validity of a series of relatedness judgement with the argument that chance is the supposedly inferring factor. However, it is by far more difficult to quantify the impact of chance. For this reason, abundant widely recognized serious research is still rejected by a evenly serious part of the linguist community**14**. It remains to be seen if the focus on less, yet measurable evidence can help in the discussion.

Only very limited amounts of items have a capacity to identify long-range relatedness directly. Even in the small list in use, we have to accept the disappearance of visible evidence due to semantic shift: for example in the concept of “dying” (‘to die’, ‘death’), the stem “-M-R-T-“/”-M-R-“/”-M-T-“ is present in many languages, even across macro families. In Germanic languages however, this stem has shifted the ‘active’ concept of death: “Murder”. It is often frustrating – but necessary – not to make use of all visible evidence – sticking to the set of rule has to be done without exception.

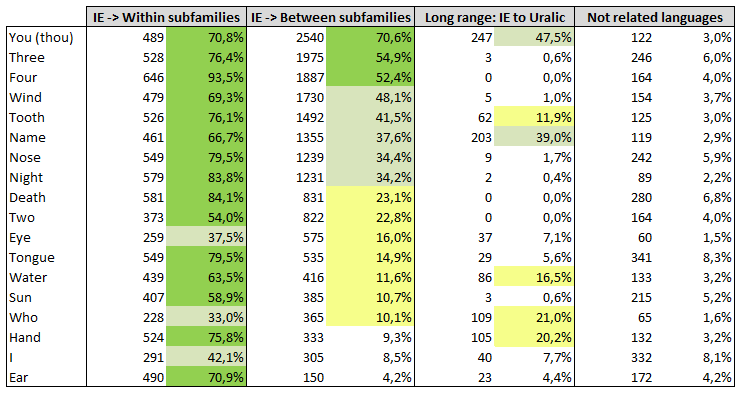
*Basic vocabulary is resistant to influence from outside*

The inferred word list happens to be very resistant to borrowing. It resists to the influence of French in English and Arabic in Iranian and Turkic languages. Exceptions have to be identified – most notably the influence of Indo-Aryan in Dravidian (three words of the list), Arabic within the Afroasiatic family and toward other language families (numerals) as well as Chinese on most Southeast Asian languages. Whenever identified, the words are excluded.

*Final word list*

The result is a set of 18 roots we illustrate in table I in the context of their performance in different time ranges and their exposure to chance.

Table 1: The selected word subset and its performance



The list reflects the percentage of words bringing more than 50% points in pairwise comparisons. The first column reflects the results for comparisons within Indo-European subfamilies (short range). The second column reflects the performance of the words in identifying relatedness between the Indo-European subfamilies (e.g. Slavic to Romance or Iranian to Germanic – medium range). Certain words still perform well at this level when other words tend to lose track of their more remote relatedness. The third column reflects comparisons of Indo-European languages with Uralic languages (long range). Words like “you (thou)”, “name”, “who” and “hand” still retain relatedness signals. The fourth column shows how the words perform among not related languages and the fifth column the values generated with random values (details about random values are addressed in §VI), quantifying the impact mere chance interference has. As we will see later in §VI, chance interference is not only useful to verify the validity of signals: it is also a powerful tool to distinguish weak relatedness signals from chance related ones.

# WORD ENCODING

To handle computer comparisons, we have to make the selected roots computer-processable and encode them accordingly. The encoding rules have to be clear and stable all across the study. Moreover, the rules’ capacity to deliver relatedness signals have to be stronger than their exposure to chance.

*Only consonants are taken into account*

Keeping only consonants – and excluding vowels and diphthongs - is a choice in the design of this study. This does not mean that we consider vowels and diphthongs play no role in comparative linguistics. In some cases, vowels resist longer to change than consonants. However, modelling vowels change leads to fewer clusters (cluster explanation bellow) than with consonants, which brings a higher exposure to chance. In certain languages, especially in ancient ones, vowels are not available. Their rapid evolution and relative instability is most remarkable in dialect differentiations. It is not a rare observation to find vowels shift within one generation**15**. Our experience hints at greater value of vowels for short-range comparisons – the value decreasing with the researched time depth.

In few cases, most notably in the classification within the Polynesian subfamily of Austronesian, which has been very conservative in its vowel systems**16**, the lack of vowel assessment leads to a misclassification within the subfamilies (e.g. Tahitian, Maori…) – these biases are certainly present – to various extend - in other language groups. They are “the price to pay” for an overall, more simple and stable system.

*Consonants are grouped into clusters*

There are many consonants in use in world languages. However, for some languages, modern and ancient, only a sound approximation is available. For this reason, we have grouped the consonants in use into 24 clusters.

Consonants share a same cluster when they differ only slightly from each other, underlying similar change rules toward other clusters. This clustering process was not computer assisted.

Table 2: The consonant clusters used in this study

|  |  |  |
| --- | --- | --- |
| **Cluster code** | **IPA code of clustered sounds** | **Examples .** |
| -B- | B | b, Cyrillic “б”, Arabic “ب” , Hindi "ब, भ", Urdu "ب بھ" |
| -C- | ts, dz, tʃ | ts, German "z" (partial), Slavic "c", Cyrillic "ц" |
| -CH- | tʃʲ | tch, German "tsch", Slavic "Č", , Kurdish "Ç", Polish "cz", Croatian/Serbian, Polish "ć", Armenian "չ", " ջ", Cyrillic "ч", Hindi "छ", Persian "ﭺ" |
| -D- | d, dˁ | d, Cyrillic “д”, Arabic "د, ض", Hindi "द" |
| -F- | f, ɸ | f, ф, ف, Afrikaans "v", Dutch "v", German "v" |
| -G- | g, ɟ | g, Cyrillic “г”, Persian, Pashto "ﮔ", Hindi "ग, घ" |
| -H- | ħ | h, Arabic "ح, ه" |
| -J- | j | Hindi "य", Arabic "ي", Slavic "j", Various languages "y" |
| -K- | k, q | k, q, Italian "ch", "qu", "cq", French "qu", Catalan "q", Galician "c", French "c" (partial), Celtic "c", Arabic "ك, ق",... |
| -KH- | x, χ, ɣ ç | Cyrillic "х", Dutch "g", German "ch", Spanish "j", Croatian/Serbian "h", Greek "χ", Gothic "X", Breton "c'h", Irish, Welsh "ch", Arabic "خ" |
| -L- | l, ɬ ʎ | l, Cyrillic “л”, ل , ल, Welsh "ll", Polish "ł” |
| -M- | m, m̥ | m, م, म |
| -N- | n, ɳ, ɳ̊, ɲ̊, ɲ, ŋ̊, ŋ, ɴ | n, Polish "ń", Arabic ن, Hindi न |
| -P- | p | p, Cyrillic “п” |
| -R- | r, ʁ, ʀ, ɾ, ɽ | r, Cyrillic "р", Arabic "غ and ر", Hindi "र" |
| -S- | s, sˁ | s, Cyrillic "с", Arabic "س, ص" |
| -SH- | ʂ, ʃ | Afrikaans "sj", German "sch", "s" before "t", French "ch", Slavic "Š", Polish "ś", Arabic "ش", Cyrillic "ш" |
| -T- | t, tˁ | t, Arabic “ت, ط " |
| -TH- | θ, ð | th, Gothic "þ", English "th", Cornish "dh, th", Celtic "th", Arabic "ث ذ", Hindi "थ, ध, ठ, ढ" |
| -V- | v, ⱱ | v, Cyrillic "в", Afrikaans "w" except after "d, k, s, t", Dutch "w", German "w" |
| -W- | W | English w, Afrikaans "w" after "d, k, s, t", Gothic "ƕ" |
| -Z- | z | z, Cyrillic "з", German "s" (partial), French "s" (between two vowels), Arabic "ز, ظ" |
| -ZH- | ʒ, ʑ, ʐ dʒ | Slavic "ž", French "j", Catalan "j", Hindi, Sanskrit "ज", Arabic “ج”, Pasto, Persian "ژ", Irish, Scottish Gaelic "dh", Polish "ź, ż", Cyrillic "ж" |
| -7- | ʕ | Arabic “ت, ط " |

An experimental, extreme simplification in eight clusters, encoding the words with numbers instead of consonants was tested in the beginning of the research**16**, delivering a more approximate but still valid classification. It illustrates how stable a classification remains even if we use an oversimplification of sound change modelling. It is by far more important to remain consistent in using the same rules for all comparisons than looking for exhaustive rules.

*Only roots matter*

When encoding the words according to the cluster system, only lexical morphemes are taken into account. Grammatical elements like nominative marks (e.g. Latin, Hittite, Lithuanian, Gothic "s" endings) or infinitive marks of verbs (Germanic "n", Slavic "t", Romance "r"...) are ignored in the cognate scoring during lexical comparison: they obviously lead to overestimation of short-range classifications within their respective subfamilies and bring an additional exposure to chance.

# Sound change modelization

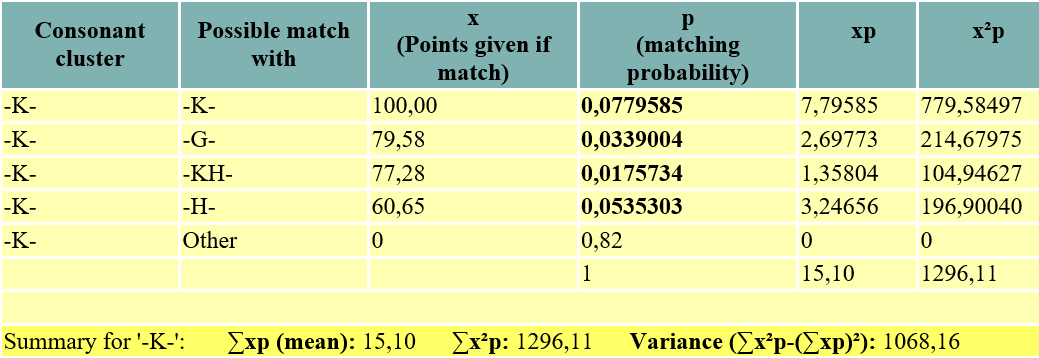
To quantify differences between sets of data, a clear and stable model has to be established. Sound change quantification is the best way to measure erosion. Sound change is very complex. To process our comparisons, we need a set of rules, which applies universally and is subject to low chance interference. Selecting these rules is the same process as for the choice of lexical items: we do not try to integrate as many items as possible, but want to use the most stable and universal ones. For this reason, we use only the part of sound change that is identified as relevant in most language families. It is not our goal to be exhaustive: we need to choose the strongest signals, carrying the lowest chance interference and process the system with the same rules all over the comparisons.

*The sound change rules used in this study are a statistically selected subset*

It is fundamental to understand that the sound correspondence rules selection process is similar to the lexical item selection process: the aim is to retain only a subset of rules, which is stable over all language families and time ranges. This objective can seem utopic and it is clear that if a subset corresponds to these criteria, it will necessarily be limited. Just as we use only an 18 words list, which seems to bring the strongest signals for relatedness measurement, we chose the items from sound change which appear to be the most stable across language families and which bring little exposure to chance. This subset is selected through the same statistical process of error reduction trying different combinations of rules and measuring their impact both on classification and on bias. The resulting, limited set of sound correspondence rules can never be compared with more elaborated sound correspondence rules used by comparatists. The purpose of their use and the methodology of their inference are completely different.

The potential of chance interference carried by the various clusters varies strongly: the number of occurrences of each consonant group in the world languages is far from constant – details of the occurrences and of their impact on chance is available on the blog**18**. An example in Table 3 shows how the rules are modelized – both as a point system and as a chance interference measurement.

Table 3: Example of sound correspondence rules used to process the comparisons



In a few cases, when languages undergo a very specific sound change, the system tends to overestimate their distance from their core-family, because it skips relatedness signals. This is most notably the case with Armenian, which has undergone a very specific sound change**19**, which therefore cannot be reflected in this study. This issue certainly affects most languages to various degrees: this is a limit to the system, which affects the stability of the results: the case of Armenian is the most obvious one, but we can assume that all languages are affected to various degrees, which has an impact on the variance of the results.

*Quantification of the sound correpondence*

Some of the clusters have a correspondence with each other. Quantifying this correspondence is necessary for the classification. Using a point system based not on intuition, but on a quantitative basis of the sound correspondence is a more convincing approach.

A quantitative approach to sound correspondence has been established by Cecil H. Brown, Eric W. Holman and Søren Wichmann in their paper “Sound Correspondences in the World’s Languages”**20**. The sound clusters are different to the ones used in this paper. For this reason, the data from “Sound Correspondences in the World’s Languages” has been aggregated to the clusters in use in this paper. After a series of tests, it has proved to be best practice (trade-off between signal strength and exposure to chance) for the starting set of rules to keep only the sound correspondences with values bigger than “3” in the original source. The values are then converted to the point system of this study as shown on the distance matrices in tables 4 and 5.

Table 4: aggregated values from “*Sound Correspondences in the World’s Languages*”**12**

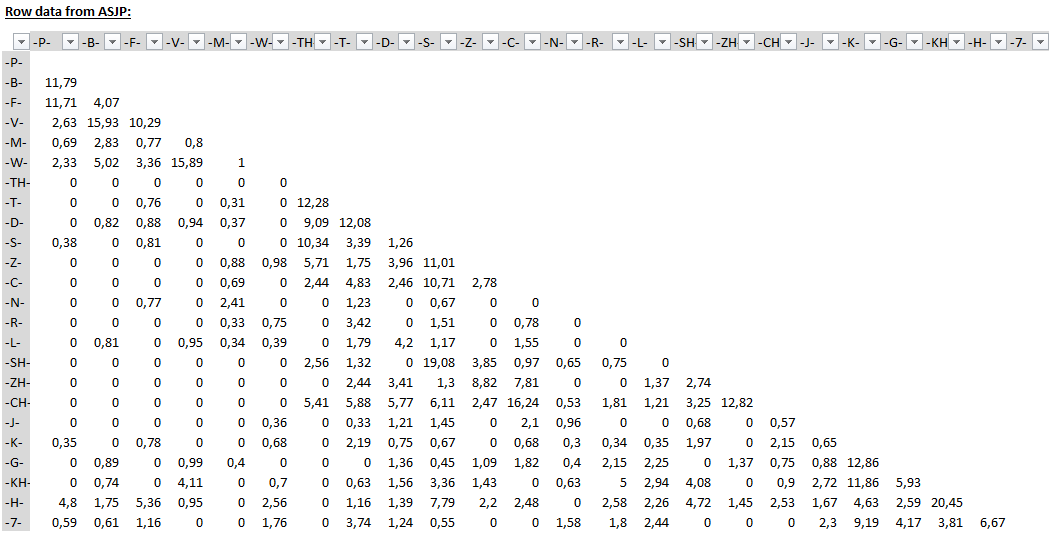
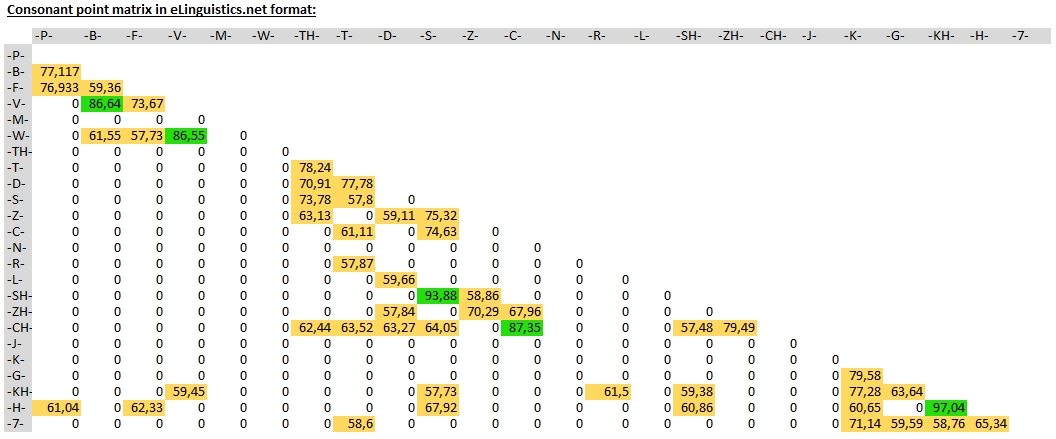


Table 5: Converted aggregated values from “*Sound Correspondences in the World’s Languages*”**12**



- Points converted to the 0-100 format, excluding weak signals –

The limitation of the number of items is not arbitrary, but underlies the same long selection process by trying and measuring the impact on known classifications as applied in the choice of the lexical items. The values in the matrix are the point given in pairwise comparisons when consonants are being compared within words.

# Pairwise comparisons

The basis for similarity judgement is the set of codified words for the compared languages. What remains visible from the languages for the computer program are the sets of 18 consonant chains as represented in table 6.

Table 6: Representation of the chain of language features used to process comparisons

|  |  |  |
| --- | --- | --- |
|  |  |  |

Processing a pairwise comparison consists of three steps at following levels:

* *Syllable to syllable level*: at this level, each syllable of the first word is compared with the syllables in the second word – points are given according to the system described in §IV – 0 corresponds to no resemblance, 100 to the same consonant.
* *Word to word level*: the results from the syllable-to-syllable comparisons are compiled together. Matches occurring in the wrong order are eliminated. The points are averaged at word level, to obtain a value from 0 to 100. Specific rules are applied to eliminate points obtained with too weak signals.
* *Language to language level*: the points obtained at word level are added and averaged to obtain a value from 0 to 100 for a comparison between two languages. This value is reversed to a distance value by subtracting it from 100. The result of a pairwise comparison between two languages is a distance between 0 (same language) and 100 (no single relatedness signal).

Unlike most quantitative approaches, the system does not calculate the proximity between languages by counting the number of cognate words in a sample, but by calculating the degree of similarity two words have. This makes a differentiation of the degree of erosion measurable. For example, the word “water”, codified [“-W-T-R-“] gets more points when compared to German “Wasser” [“-V-S-R-”] than to Russian “вода” (voda) [“-V-D-“]. This system does not guarantee that the proximity between two words will always follow the actual proximity between languages – example like English “night” [“-N-T-“] which consonants better match Italian “notte” [“-N-T-”] than German “Nacht” [“-N-KH-T-“] shows the variations happening in the details. The distance value at language level is 48 for English to Italian and 31 for English to German, which reflects better the degree of difference between these languages than it does with comparison of one word only.

Distance values are relative to each other: A distance between two languages can only be interpreted in the context of their respective distances to other languages.

The distance results obtained with this system are available online at the author’s blog**1** since February 2013 and have aroused a broad interest, resulting in numerous comments and suggestions for improvements, especially with regard to the choice of the roots and their right encoding in various languages. It has also revealed in practice that, even if the single values of pairwise comparisons most of the time correspond to what the users expect, there are variations that lead to surprising results. The causes are numerous and reflect important limits:

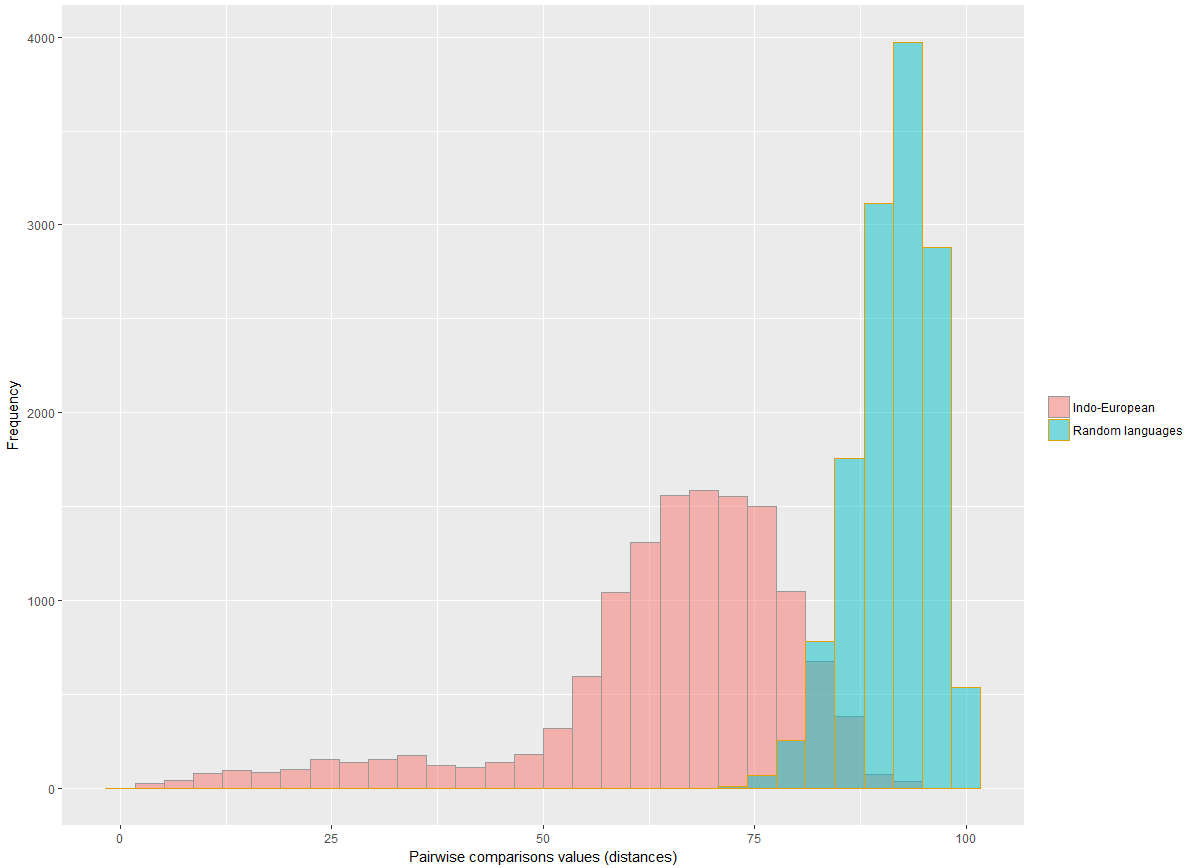
* The system assumes a constant rate of evolution, which does not correspond to the reality. Languages evolve at different speeds**21** in time and place.
* We compare languages in different states. Even if the root list in use is resistant to massive vocabulary loans, other aspects in the history of languages influence their evolution artificially. Modern languages are often codified compromises between dialects**22**.
* In medium and long-range comparisons, the variations can be due to chance exposure, which – even if we strive to minimize it – still impacts the values.
* In short-range comparisons, the number of differences is small, giving each of the difference an overestimated weight. This phenomenon is most clearly visible in pairwise comparisons within the Slavic subfamily. If the aim is to be more accurate on short-range comparisons, more lexical items are needed.

As we will see in the further steps of the project (§VII-IX), the most interesting values are not the single comparisons but the complete dataset of all comparisons, their distribution and their clusters. The statistical significance of the results grows with the amounts of data. In the mass comparisons, biases, which in some cases heavily affect single pairwise comparisons, tend to compensate each other, leading to much more interesting results.

# Statistical context

The central issue across this study is to determine if an observation can be due to chance, and if so, to which extent. The null hypothesis is that the results are normally distributed around a mean of 91.45 and a standard deviation of 4.71. An observation of the distribution of all pairwise comparisons within the Indo-European family, plotted together with a distribution of an equally sized quantity of pairwise comparisons between randomly generated languages reflects the challenge we face to interpret the results.

Table 7: Distribution of pairwise comparisons between IE languages vs. between random languages



The results clearly show that the distribution of the Indo-European pairwise comparisons do not follow a normal distribution (the null hypothesis is rejected and the results are not due to chance). Results of pairwise comparisons between randomly generated languages reflect the impact of chance. These results are normally(?) distributed. The distribution of randomly generated results is very similar to the distribution of pairwise comparisons between not related languages.

The challenge is that pairwise comparisons with a distance result above 65 can reflect a relatedness signal with a relevant probability to be due to chance. At this level, we are in the scope of the most interesting results, as they could reflect very old relationships. However, the single results have to be interpreted with great caution: a seemingly steady result, with a confidence interval of 99% is one of hundreds of results – as there are many potential pairwise comparisons revealing supposedly interesting findings. Within the mass, it is always possible to find a “convenient” result – we have to bring objectivity here.

*Calculation the p-Value of pairwise comparisons*

Since we are using a clear and systematic set of rules to quantify the distance between two languages, we are also in a position to infer the expected value and standard deviation specific to a comparison. These inferred values are specific to each language and comparison and differ – sometime strongly – from the observed values above.

Not all languages use the same consonants to the same extent or have the same number of syllables in the words. With a constant number of words in the comparison, the expected value and expected standard deviation is specific to each language. Since we not always have 18 words for all languages (ancient languages for which the words are not available or modern languages where known borrowing are excluded), the expected values and expected standard deviations will also vary according to which language a language is compared.

*< MISSING: STEPS OF THE P-VALUE CALCULATION WITH FORMULAS > p-Value is used for exploratory data analysis + positioning languages with too little material for automatic clustering/inferring. Doesn’t affect research in dating and exploring long ranges.*

*Bootstrapping & Monte Carlo Markov Chain*

As mentioned before, the real value of the data lies in the mass comparisons, their clustering and their distributions. At these levels, advanced techniques are necessary to cope with the influence of chance on results. We will handle them in their respective contexts in §VII, VIII and IX. As we will see, these techniques are not a mere help to validate inferred results: they are in the very core part of the system.

# Clustering – phylogenetic classification

With the ability to produce distances between languages in pairwise comparisons, we can produce a large dataset with comparisons between all possible pairs of languages. To obtain meaningful insights we have to identify clusters of similar datasets. The most obvious way of visualizing clusters from a large amount of distance data is to produce phylogenetic trees. The process of obtaining a phylogeny from distance data is very straightforward. The data have to be organized in a distance matrix. To process phylogenies from distance matrices, many programs are available. For this study, we choose to use MEGA**23** and the APE-Package**24** of R**25**.

*Choosing the right algorithm*

The phylogenetic tools offer different algorithms, each of which have strengths and weaknesses. A basic question, which has to be answered to choose the right algorithm, is whether the data is ultrametric or not. Because we assume a constant rate of evolution and because the distance to a common ancestor is equal for all ist descendants at a given time, the data we use can be considered as ultrametric. In this case the simple UPGMA algorithms delivers good result. The Neighbor Joining and Minimum Evolution algorithms deliver phylogenies, which are very similar to UPGMA and are considered more reliable. We use all of these algorithms, both in MEGA and in R/APE to validate the stability of the results.

When the raw phylogenetic tree is produced, we have to make sure we identify nodes and branches, which we can rely on. Nodes and branches which vary when we change the algorithm or when we add or subtract datasets are not stable. The classical method to measure node stability, bootstrapping the phylogenetic tree, cannot apply here, because the distance values we use in comparison are obtained in a complex process: it does not make sense to resample the matrix. We have to resample the data which the pairwise comparisons process**26**.

*First validation challenge – validating macro families and links between them*

The most interesting nodes and branches in the trees are the ones limiting and connecting macro families. The distance values at which level macro families get connected is high and their connection relies on low statistical confidence levels. The intuition is that we have to be doubtful about these connections. But how to determine more precisely where?

To address this issue, “random languages” are generated and processed with the comparisons. The same number of random languages as actual languages is included in the distance matrix. We then use the position and values of the performance of random languages to determine which nodes can be validated.

The process of generating random languages is complex: the comparisons of these languages between each other has to follow the same distribution as the comparisons between not related languages.

< Monte Carlo Markov Chain? >

*Second validation challenge – validating nodes and branches between subfamilies*

The connection of nodes between subfamilies are often at a distance level, which cannot be validated or invalidated through resampling. The uncertainty at this level is simply due to the relative instability of the results. We have is to consider and process these results in the frame of their variances and accept or reject drawn connections accordingly.

*From raw tree to validated tree*

The raw tree includes all branches, including random languages. To produce the final tree, we first apply the correction obtained through resampling: wherever a random language gets classified between branches of real languages, we consider that there can be no connection between languages separated by random languages. This process eliminates irrelevant connections between macro families and validates relevant one.

The second correction that has to be applied is to eliminate nodes which cannot be validated due to high variance. The branches are connected to the next upper nodes. This process reduces the granularity of the classification.

The third necessary modification is the repositioning of the ancient languages. During the clustering process, the system does not take the period in which the language was spoken into account. Latin, Avestan, Old Greek are all processed the same way as modern languages. To correct this, the position of the branch connection has to be placed higher in the tree. This process is done according to the data inferred in the dating process (§ VIII).

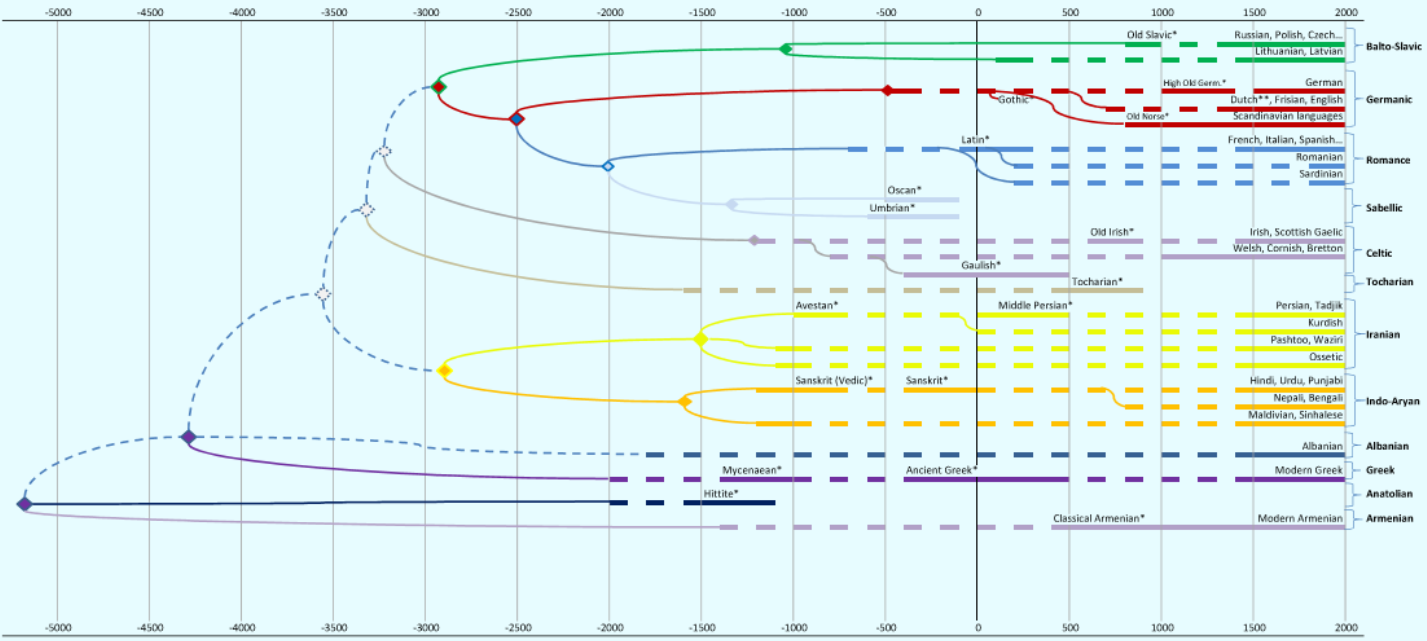
*Special case: languages with fewer than 15 available words*

Languages with fewer words, especially the ones with very few known words like Phrygian, Mycenaean, Oscan, Hurrian, etc. bias the classification because they tend to find similarities simultaneously with several language families. These language are processed manually on the basis of the distribution of their distances to other languages: if a clear group of languages seems to be the next related one, the language is placed near to it manually.

|  |  |
| --- | --- |
| Raw tree | Processed tree |
|  |  |

*Validating and stabilizing the results*

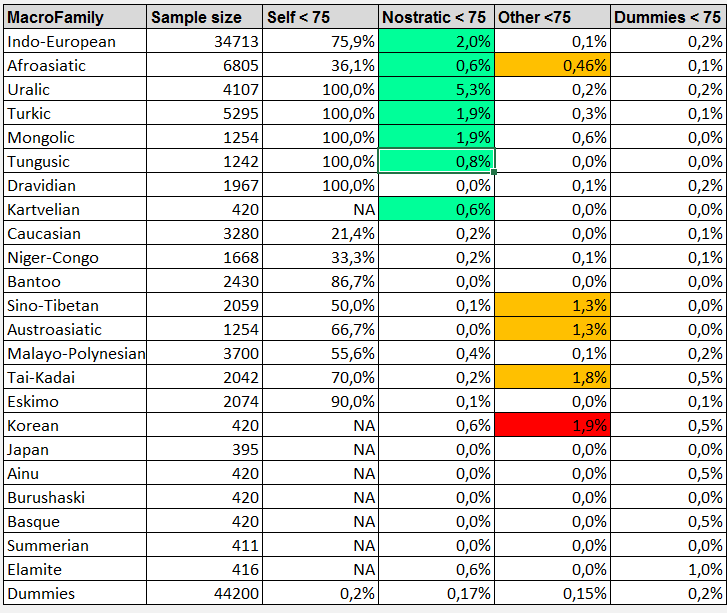
# Dating nodes



Date comparison to ASJP/Ringe/Berkerley + Train/validate!

# Exploring long range

Some single comparisons and phylogenetic classifications hint at connections between languages across macro families, yet without sufficient statistical evidence. Using the complete dataset, including the random languages, we have the possibility to analyze distributions according to fixed criteria. We know that distance results below 75 tend to infer relatedness. However, we also know that results above 65 have a non-negligible probability to be due to chance. To identify possible relationships between macro families, we analyze their performance in a specific context.



Position in Nostratic/Eurasiatic research + attempt for Caucasian/Austric

# Conclusion

***The system gives a tool to watch in the past. As a lamp directed in the dark of the past it delivers an unsharp view (variations) especially on single items. But the light it brings is useful for periods of several millennia, becomes weaker and weak.***

***The age of the separation of language families and subfamilies is estimated as a wider range. The several ranges overlap in many cases, so that this system doesn’t answer the question of nodes’s ages with certainty. But its classification power can help in the consideration of other material (archeological, genetic…) With the common view on Eurasiatic and Nostratic, we reach the most interesting results but also the time limit in which a***

***The fact that this method confirms only overlapping hypotheses can be seen as a strength. It gives light as far as into the Eurasiatic and Nostratic hypotheses, yet without any clue for not widely accepted parts of it.***

***Disadvantages of the system***

***By applying the same rules across the system, we miss potential signals. This aspect is the necessary sacrifice we need to stick to an automated system and yield the advantages from it.***

REFERENCES:

**1** http://www.eLinguistics.net/

**2** Reference for established language family classifications: https://www.ethnologue.com/browse/families.

**3** Swadesh, Morris (1955). Towards greater accuracy in lexicostatistic dating'. International Journal of American Linguistics 21, p. 121–137.

**4** Swadesh, Morris. (1971). The Origin and Diversification of Language. Ed. post mortem by Joel Sherzer. Chicago: Aldine. ISBN 0-202-01001-5. p. 283.

**5** List, Johann-Mattis & Cysouw, Michael & Forkel, Robert (eds.) 2016. Concepticon. Jena: Max Planck Institute for the Science of Human History. (Available online at http://concepticon.clld.org, Accessed on 2018-01-02.)

**6** Wichmann, Søren, Eric W. Holman, and Cecil H. Brown (eds.). 2016. The ASJP Database (version 17).

**7** Machine Learning Using R, A Comprehensive Guide to Machine Learning, Karthik Ramasubramanian, Abhishek Singh, p. 222

**8** Simon Rogers,Mark Girolami, A First Course in Machine Learning, Second Edition, CRC Press, 2017

**9**  Mark Pagel, Quentin D. Atkinson, Andreea S. Calude, and Andrew Meadea, Ultraconserved words point to deep language ancestry across Eurasia, PNAS, vol. 110 no. 21, 8471–8476, May 21, 2013

**10** V.M. Illych-Svitych / В.М. Иллич-Свитыч - Опыт сравнения ностратических языков – ISBN: 5-354-00173-0 (1971)

**11** Aharon Dolgopolsky, The Nostratic Macrofamily and linguistic Palaeontology, Tje McDonald Institute for Archaeological Research, 1998.

**12** Josef H. Greenberg, Indo-European and its Closest Relatives: The Eurasiatic Language Family Volume II: Lexicon. Stanford, CA: Stanford Univ Press; 2002.

**13** Allan R. Bomhard. A comprehensive introduction to Nostratic comparative linguistics, Volume 2, third edition 2018.

**14** Iaroslav Lebedynsky, LES INDO-EUROPÉENS Faits, débats, solutions – Editions Errance, ISBN : 978-2-87772-564-4, p. 81

**15** Trask, Introduction to Historical Linguistics, p ?

**16** Variation in Indigenous Minority Languages, p. 146, James N. Stanford, Dennis R. Preston, John Benjamins Publishing Company

**17** http://elinguistics.net/Comparative\_Linguistics/Cluster\_Evolutionary\_Tree.html

**18** <http://elinguistics.net/ConsonantFrequency.html>

**19** Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics, p. 116, Hans Henrich Hock, Brian D. Joseph, Editor: [Walter de Gruyter](https://books.google.at/url?id=IsYkilw7Q-oC&pg=PA116&q=http://www.degruyter.com&clientid=ca-print-degruyter&linkid=1&usg=AFQjCNEao0N00npWUevz-V8xqcxwJLntrg&source=gbs_pub_info_r)

**20** Cecil H. Brown, Eric W. Holman and Søren Wichmann, “Sound Correspondences in the World’s Languages” - https://muse.jhu.edu/article/503022/pdf

**21** Evolution rate of languages -> Reference in Trask?

**22** Official languages artificial, compromise between dialects -> Reference in Trask?

**23** MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger datasets (submitted). Kumar S, Stecher G, and Tamura K (2015)

**24** Paradis E., Claude J. & Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290

**25** R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

**26** Josef Finkelstein Confidence Limits Using on Phylogenies: an Approach Using the bootstrap, Evolution 39(4), 1985, pp. 783-791