The Distribution of Data in Word Lists and its Impact on the Subgrouping of Languages

Hans J. Holm

Hannover
HJJHolm@web.de

Abstract. This work reveals the **reason** for the bias in the separation levels computed for natural languages with only a small amount of residues; as opposed to stochastically normal distributed test cases like those presented in Holm (2007a). It is shown how these biased data can be correctly projected to true separation levels. The **result** is a partly new chain of separation for the main Indo-European branches that fits well to the grammatical facts, as well as to their geographical distribution. In particular it strongly demonstrates that the Anatolian languages did not part as first ones and thereby refutes the Indo-Hittite hypothesis.

1 Aim and general situation

This analysis of the distribution of wordlists is the - up to now - **latest step in quantitative methods** to infer the genealogical subgrouping of languages.

Traditional historical linguists use to look upon such methods with **suspicion**, because they argue that only those agreements can decide the question, which are supposed to stem exclusively from their direct ancestor, the so-called '**common innovations**', or synapomorphies, in biological terminology (Hennig 1966). But, this seemingly perfect concept has in over a hundred years of research brought about **anything but agreement on** even a minimum of groupings (e.g. those in Hamp 2005). There is no grouping, which is not debated in one or more ways. 'Lumpers' and 'splitters' are at work, as with nearly all proposed language families.

Quantitative attempts (cf. Holm 2005, updated 2007, for an overview) must in most cases be criticized for two reasons: First, many of them hold on the mechanical rate (or "clock") assumption for linguistic changes (what is not our focus here). Second, mathematicians, biologists, and even some linguists retreat to the too loose view that the amount of agreements is a direct measure of relatedness. Elsewhere I have demonstrated that this assumption is erroneous because these researchers miss the fact that the amount of shared agreements is a **surface phenomenon,** the "proportionality trap" (cf. Holm 2003; Swofford 1996).

2 The Bias

2.1 Recapitulation: What is the Proportionality Trap?

Definition. What are we talking about? If a "mother" language splits into two daughter languages, these are called "genealogically related".

These two daughter languages differ from each other as well as from their parent language(s), because languages **change**,

- **independently** (!) from each other,
- by new socio-psychological impacts in different (!) **irregular** amounts in history. It is therefore non-deterministic in that the next state of the environment is partially but not fully determined by the previous state. Least, it is
- **irreversible**, because, when a feature is changed, it will normally never reappear. Because of these properties we have to regard linguistic change mathematically as a **stochastic process** with draws without replacement.

Though, the changes are unforeseeable and neither projectable into the past nor the future in a glottochronological sense, there exists a **tool well known in statistics,** what any mathematician would immediately recognize as the **hypergeometric distribution**. In word lists, we in fact have all **four parameters** (cf. Fig.1) needed as follows:

- The amount of **inherited features 'k_i' and 'k_j'** (elsewhere residues, cognates, symplesiomorphies) regarded as preserved from the common ancestor of any two languages L_i and L_j;

- the amount of shared **agreements '$a_{i,j}$'** between them;

- the number N of their common features at the time of separation (the **universe), not visible** in the surface structure of the data.

Exactly this universe N we are seeking for, because it represents the **underlying structure,** the amount of features, which must have been present in both languages at the era of their separation. And again any mathematician has the solution: This “**separation level” N** for each pair of branches can be inferred by the 2nd momentum of the hypergeometric, the maximum likelihood estimator, transposed to

$$\hat{}N = k_i\, k_j / a_{i,j}. \qquad (1)$$

Since changes can only lower the number of common features, a higher separation level must lie earlier in time, and thus we can obtain a **chain of separation** of a family of languages.

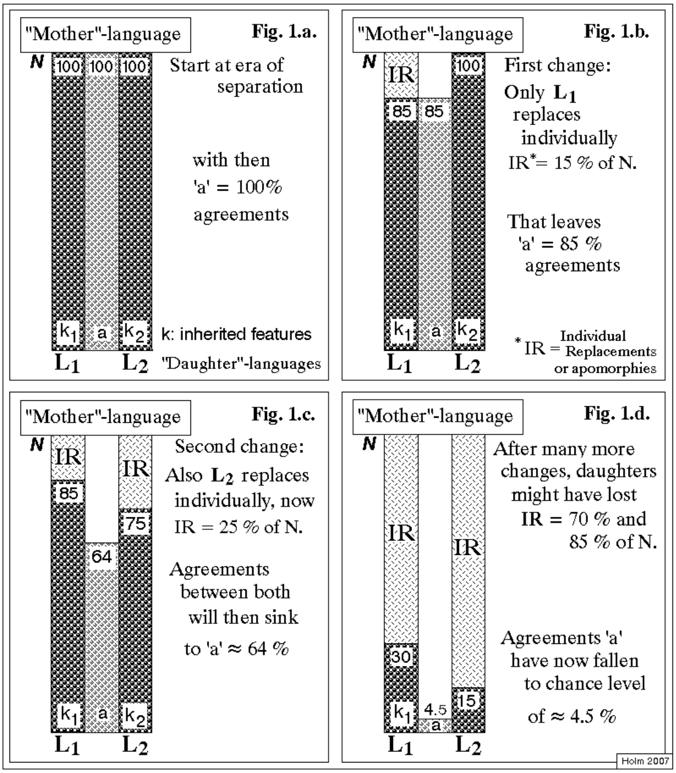

"Mother"-language
Fig. 1.a.
N
100
100
100
Start at era of separation
with then 'a' = 100% agreements
k1 a k2
k: inherited features
L1 L2
"Daughter"-languages
"Mother"-language
Fig. 1.b.
N
IR
100
85
85
First change: Only L1 replaces individually IR*= 15 % of N.
That leaves 'a' = 85 % agreements
k1 a k2
* IR = Individual Replacements or apomorphies
L1 L2
"Mother"-language
Fig. 1.c.
N
IR
IR
85
75
64
Second change: Also L2 replaces individually, now IR = 25 % of N.
Agreements between both will then sink to 'a' ≈ 64 %
k1 a k2
L1 L2
"Mother"-language
Fig. 1.d.
N
IR
IR
After many more changes, daughters might have lost IR = 70 % and 85 % of N.
Agreements 'a' have now fallen to chance level of ≈ 4.5 %
30
15
4.5
k1 a
L1 L2
Holm 2007

Fig. 1: Different agreements, same relationship

2.2 Applications up to now

These insights have been applied to the problem of genealogical subgrouping of languages. The first one to propose and apply this method was the British mathematician D.G. **Kendall** (1950) with the Indo-European data of Walde/Pokorny (1926-32). It has then independently been extensively applied to the data of the improved dictionary of **Pokorny** (1959) by this author (Holm 2000, passim).

The **results** seemed to be convincing, in particular for the North-Western group, and also for the relation of Greek and the Indo-Iranian group. The late separations of Albanian, Armenian, and Hittite could well have been founded in their central position and therefore did **not appear suspicious**.

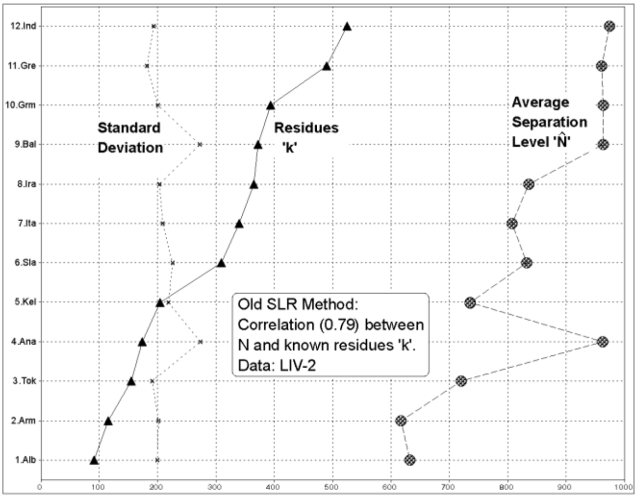
Only when in a further application to **Mixe-Zoquean** data (Cysouw et al., 2006) a resembling observation occurred that only languages with few IE residues appeared to separate late, a systematic bias could be suspected. Cysouw et al. **discarded** the SLR-method, because their results contradicted the subgrouping of Mixe-Zoquean as inferred by traditional methods of two historical linguists (which in fact did not completely agree with each other). In a presentation, published in the web as CysouWIP.pdf (09.11.2004) he stated that the "unbalanced amount of available data distorts the estimates". But, having understood the basics explained above, this could **not logically be the true reason**.

Meanwhile I had started to evaluate the most modern and acknowledged Indo-European dictionary, the "Lexikon der indogermanischen Verben" (Rix et al. 2002, second edition, henceforth **LIV-2**. I am very obliged to the authors for sending me the digitalized version, which in fact only enabled me to quantify the contents in acceptable time). The reasons for this tremendous undertaking were:

- the commonplace (though seldom mentioned) in linguistics that **verbs are much lesser borrowed** than nouns, what is not taken into account by any quantitative work up to now. Everybody can easily find examples, e.g. in German, "Ich arbeite mit dem Computer" but never, "Ich worke mit dem Rechner"; or in English, "The Iraq war seemed to be a blitzkrieg" rather than " ... schien a speedy victory zu werden".
- the more trustworthy combined work of a team at an established department of Indo-European under the supervision of a **professional Indo-Europeanist** should guarantee a very high standard, moreover in the

second edition.

- Compared with the in many parts outdated Pokorny, we have now much **better knowledge of the Anatolian** and Tocharian languages.



Old SLR Method: Correlation (0.79) between N and known residues 'k'. Data: LIV-2

Fig. 2: False correlation k-N from LIV-2 list.

Nevertheless, these much better data, **not suspicious of poor knowledge**, displayed the **same bias** as the other ones, as can be seen in fig.2 presenting the correlation between the residues 'k' and their resulting '^Ns' in falling order.

Thus we have a problem. The reason for this bias, opposite to Cysouw et al., could not lie in a poor knowledge of the data, nor could it lie in the algorithm, since that is mathematically reliable, as I have additionally tested in hundreds of random cases, some of which published in Holm (2007a).

Consequently, the reason had to be found in the word lists alone, the properties of which we will have to inspect with more scrutiny:

3 The Reason

3.1 Revisiting the properties of wordlists

I have already intensively investigated the effects of **scatter** on the subgrouping problem as well as its handling. To avoid excessive scatter in hypergeometric estimates, textbooks as well as a hundred tests suggest that the sum of residues 'k' should at least amount to 90 % of the universe 'N', and a single 'k' must not fall below 20 %. Within these limits, I have presented several cases with complicated subgrouping configurations in Holm (2007a) and demonstrated that and how this problem can be solved. But, since the LIV-2 database is big enough to guarantee a low scatter, there must be something else, overlooked up to now.

A first hint has already been given by D.G. Kendall (1950:41) who noticed that,

"One must, however, assume that along a given segment of a given line of descent the **chance of survival** is the same for every root exposed to risk, and one must also assume that the several roots are exposed to risk **independently**".

The latter condition is the easier part, since linguists would agree that changes in the lexicon or grammar occur independently of each other. (The so-called push-and-pull chains are mainly a phonetic symptom and of lesser interest here). The real problem is the first condition, since the **chance of survival is not at all the same** for any feature, and every word has its own history.

For our purpose, there is no reason here to deal with the **reasons** for these changes in detail. We may just mention that mainstream opinion seems to be that the chance of survival is higher with the frequency of usage. This opinion has in turn led to the compiling of so-called basic word lists consisting of one to two hundred lexemes assumed to be the most stable ones.

Could the reason for the observed bias perhaps be found in a **distribution that contradicts** the conditions of the hypergeometric, and perhaps other quantitative approaches, too?

Such distributions have already been described by Pareto, and by the so-called **Zipf's Law**, originally saying that the product of the rank R of a word multiplied by its squared frequency F is constant (Zipf, 1965). Because this law has been disputed, and meanwhile changed in several ways, we are left to analyze the data ourselves:

3.2 Detecting the distribution

To find out the frequency distribution, we take our source list as available in the mentioned LIV-2.

The 1195 reconstructed verbal roots have to be entered into the rows of some spreadsheet as characters, while the **columns** contain the 12 branches of Indo-European. The cells or cross-fields then contain a 1, if a cognate of the root is identified, and a zero, if not so. A smaller problem must be solved here: In the LIV-2, as in other work, we encounter **questionable** reconstructions. We could now simply count the observations, not regarding their different degree of reliability. I preferred to achieve a higher level of validity by **weighting**, e.g. by allocating 0.5 to reconstructions with a question mark.

We let now sum up the cross totals (of the digits) into a new column, containing then the frequency for every row. Next we **sort** the whole dataset after these frequencies. Since the lowest observation is 1 (the word occurs in one branch only), while the highest is 12 (the verb occurs in all branches), we get twelve different blocks or slices, one for each frequency. After counting out every language per slice, and entering these numbers into a new table, we obtain the **plot** of fig.3.

(Note in proof: Since we have discrete data the plot should be a histogram, but that would be nearly unreadable).

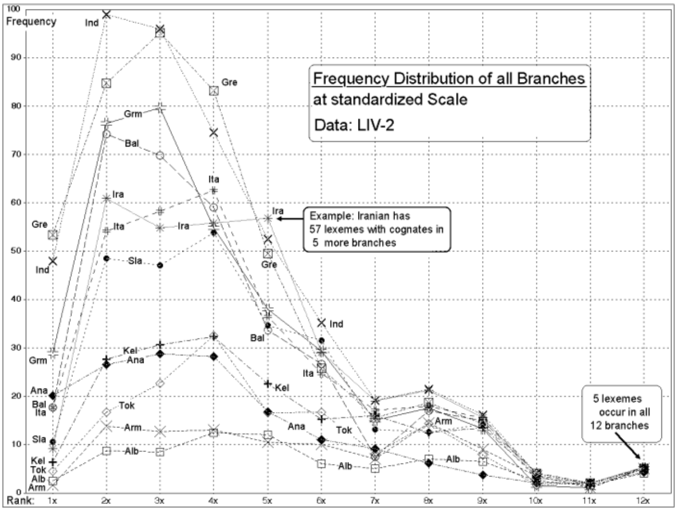
3.3 Analysis of the distribution

Immediately we observe to the right hand the few verbs which occur in many languages, growing up to the left with the many verbs occurring in fewer languages, breaking down to the special case of verbs occurring in one language only (where is always a problem of prove).

Mathematicians would call these distributions 'skewed to the left'. But rather as to identify the Zipf or other distributions, we have to look for the reason of the false correlation between these curves and the **bias** with the smaller represented languages.

Where are the connections with our formula? ^N depends on the product of the residues 'k' of any language, as represented here as the area below their curve. This is then divided by their agreements 'a'. But where are these **agreements** in this graph? In fact these are **represented by the frequencies**. And there we are: The more to the right hand, the higher the agreements per residue. This is the deciding point: We observe that the **smaller**

the sum of residues of a branch**, the higher is the proportion of agreements**, ending in a false lower separation level.
So far we have **located the problem**. But we are still far from a solution.



Frequency Distribution of all Branches
at standardized Scale
Data: LIV-2

Example: Iranian has
57 lexemes with cognates in
5 more branches

5 lexemes
occur in all
12 branches

Fig.3: All frequencies of LIV-2 list

4 The Solution

4.1 Operationalization

Looking at the suspected correlation in fig.2, one might simply try to correct the 'k's according to their observable dependence upon the sum of estimations ^N. But this would be a superficial "milkmaids" **miscalculation**, because the underlying reasons or conditions are not taken into account. These **conditions** are:

- We may only employ data with the **same chance of being changed**, what excludes to work with the total numbers, and rather limits us to the data of one single block;
- On the other hand, we have at least to employ the proportions mentioned above to avoid unacceptable **scatter**.

To observe these two **conditions**, turns out to be an extremely difficult task, because they **contradict** each other, as all tests with word lists have confirmed.

Which of the **three obvious options** would be the best?

a) First, we can select the single block with the **highest amount of words**. This would allow a simple calculation with the additional advantage of relatively low spread. Regrettably, these words with their low frequency are just those with the highest risk of being changed. Additionally they represent the part of the data with the lowest, and often unusable proportion of agreements, ending in too high or early splits, which - as tests confirmed - can exceed the universe by the double. Employing a group of this kind, e.g. blocks 2 to 5, only reduces, but not eliminates this problem. So, this option leads to wrong results.

b) The contrary would not be much better: To select the 100 to 200 words with **highest frequency** (as suggested by Morris Swadesh since the early fifties under the assumption of high stability). Here we would encounter too insignificant differences, with high scatter, because the slightest error in the few observed replacements would result in wrong estimates. Additionally, we would get the not representative portion with high agreements, resulting in too late separations. Of course this is not acceptable, either.

c) Thus we are left with the most complicated option: Since we may only compare datasets where all characters have the same risk of being replaced, we will have to compute the separation levels for **every slice independently**. By the possibility of averaging the results, we earn the additional advantage to rule out scatter to some degree.

4.2 Implementation

This solution would require (11·12)/2 = 66 per slice; times 12 slices are 792 computations. Only minor relieves are allowed: We leave off slice one, for it yields too insignificant values, anyway. Further, we combine slices with nearly equal frequencies, like 7 & 8, as well as 10, 11 and 12, what leaves us with 8 slices or **528 computations**. We thereby get a sufficient amount of preliminary Separation Levels for every pairing, as to **rule out the scatter**. Thus we do not need the median for this purpose, which is not recommendable either, because it conceals the differently skewed amounts. Outliers can additionally be avoided to some degree by leaving off those cells with agreements **below a = 5**, what is an absolute threshold for the hypergeometric, anyway. The achieved means of '^N' are then entered into the final matrix.

5 The Results

5.1 From the final matrix to new subgrouping.

There are many ways of evaluating such kind of matrices.

I do not suggest reconstructing a tree by any of the older hierarchical agglomerative **cluster analysis** methods, which start from one taxon (or single branch) only and then sequentially uphill, and which further rely too much on the concrete values, not regarding the still susceptible scatter. I will further not go into details of my former proceedings, where I worked with a huge database of the Pokorny and consequently much lower scatter.

I rather suggest proceeding on a **broad front**, starting to combine every branch with its next neighbors, detectable by the lowest SL, equaling the latest separation. This could be done by eyeball or my Bx method described in Holm (2005:640). In short, this is a measure of close historical development, which avoids possible influences from the last neighbor. As soon as we have combined the last neighbors, we proceed in the same way. We then can regard the **higher cross-field values as scatter** and work with the arithmetic mean, but again, only for the next step, not for the whole line. Of course the **most reliable** values, with lowest scatter, are those branches owning the highest retention base 'k'.

In the end we arrive at the separation levels '^N' for every pairing and the final subgrouping - as a **tree**. By the way, the above-mentioned Bx-values are in particular helpful for "**flattening**" the graph, which naturally is only ordered in the one direction of descent, but might represent different clusters or circles in real geography, what is not displayable in a two-dimensional graph.

5.2 Discussion

Linguistically, the outcome clearly refutes the Indo-Hittite hypothesis, which holds that Anatolian has left the body of pre-Indo-European first.
There remain sources of bias beyond the reach of mathematicians:

- In the data matrix appears an extremely early split between Anatolian on the one side and Baltic on the other hand. I have handled this as scatter, what must not necessarily be the case. It could rather arise from gaps reflecting different cultural background: A hunter and gatherer community as to be found in the Baltics would adopt features from other semantic fields than one in the highly developed culture areas in Anatolia. Hence replacements between both would not agree and come out as too few ones. Consequently, if we would in some way be able to regard these gaps, Anatolian would have split off even later.

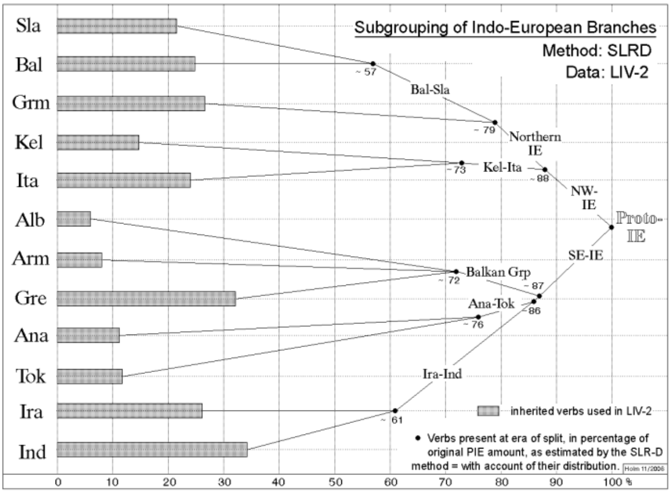


Fig. 4: New SLRD-based subgrouping of main IE branches

- Second, there will remain preferences and bias in **research** itself. Not the pure amount of data, but different scrutiny, looser or tighter measures would result in false amounts of agreements 'a'. Nevertheless, we might detect such shortcomings by otherwise unmotivated peaks or dents in the curves.
- Third, the relative **location** of a language may affect the results. A central position with long lasting close neighborhood ("Sprachbund") leads to different forms of concealed borrowing, simulating a false closer relationship, whereas a peripheral location ("Saumlage") often causes lower contacts, thus leading to very conservative habits and false earlier separation (cf. e.g. Pennsylvanian Dutch).

5.3 Conclusion

Finally, we have succeeded in finding the reason for the bias caused by different distributions in word lists, and more: We have upgraded the SLR-method to one considering the (D)istribution, which I propose to name SLRD now. Though the SLRD can solve subgrouping problems of language families with as sufficient amount of data, it will encounter difficulties with immense scatter with looser families and poorer databases like perhaps "Nostratic". Nevertheless, whatever other method is used, neglecting the results of this study, in particular working with simple agreements or neglecting the impact of the particular distribution, must lead to wrong results, what can of course always be camouflaged by accidentally satisfied conditions.

6 References

CYSOUW, M., WICHMANN, S. and KAMHOLZ, D. (2006): A critique of the separation base method for genealogical subgrouping, with data from Mixe-Zoquean. *Journal of Quantitative Linguistics*, 13 (2-3), 225-264.

EMBLETON, S.M. (1986). *Statistics in historical linguistics* [Quantitative Linguistics 30], Bochum: Brockmeyer.

Grzybek, P., and R. Köhler (Eds). (2007). *Exact Methods in the Study of Language and Text* [Quantitative Linguistics 62], Berlin: de Gruyter.

HOLM, H.J. (2000): Genealogy of the Main Indo-European Branches Applying the Separation Base Method. *Journal of Quantitative Linguistics*, 7-2, 73-95.

HOLM, H.J. (2003): The proportionality trap; or: What is wrong with lexicostatistics? *Indogermanische Forschungen* 108, 38-46.

HOLM, H.J. (2007a): Requirements and Limits of the Separation Level Recovery Method in Language Subgrouping. In: GRZYBEK, P. and KÖHLER, R. (eds), Viribus Quantitatis. *Exact Methods in the study of Language and Text. Festschrift Gabriel Altmann* zum 75. Geburtstag. Quantitative Linguistics 62. De Gruyter, Berlin.

HOLM, H.J. (to appear): The new Arboretum of Indo-European 'Trees'. *Journal of Quantitative Linguistics*.

KENDALL, D.G. (1950): Discussion following Ross, A.S.C., Philological Probability Problems. In: *Journal of the Royal Statistical Society*, Ser. B 12, p. 49.

POKORNY, J. (1959): *Indogermanisches etymologisches Wörterbuch*. Bern: Francke.

RIX, H., KÜMMEL, M., ZEHNDER, Th., LIPP, R. and SCHIRMER, B.

(2001): *Lexikon der indogermanischen Verben.* Die Wurzeln und ihre Primärstammbildungen. 2. Aufl. Wiesbaden: Reichert.

WALDE, A., and J. Pokorny (Ed). (1926-1932): *Vergleichendes Wörterbuch der indogermanischen Sprachen*. Berlin: de Gruyter.

ZIPF, G.K. (1965): *The psycho-biology of language. An introduction to dynamic philology*. Cambridge MA: MIT Press.

Keywords

INDO-EUROPEAN, SUBGROUPING, PHYLOGENY, LANGUAGES, RELATIONSHIP, LEXICOSTATISTICS.