# Beautiful Trees on Unstable Ground *

Hans Geisler, Johann-Mattis List

2009/09/25

While lexicostatistics and glottochronology were believed to be dead for a long time, the integration of stochastic methods taken from genetics has initiated an unexpected revival of these scorned disciplines. The proponents of these 'new quantitative methods' in historical linguistics claim that the procedures are relatively robust regarding errors in the data (wrong cognate judgments, undetected borrowings or wrong translations). In order to check this claim, we have investigated the differences and errors in two large lexicostatistical datasets and tested their influence on the topologies of computed family trees. Our results show clearly that the shortcomings of lexicostatistics and glottochronology have not been overcome by these new computation methods: the main problems of lexicostatistics and glottochronology, the translation of basic concepts into individual languages and the execution of cognate judgments, are still so grave that no reliable results can be drawn from this methods.

## 1 Lexicostatistics

### 1.1 Basic Assumptions of Lexicostatistics

The accounts on the key assumptions of lexicostatistics given in Arapov & Cherc (1983:17-20), Gudschinsky (1956[1964[:613) and Sankoff (1969:2f) differ slightly in some respects. We can summarize the core of lexicostatistical theory in the following two basic assumptions:

1. The lexicon of every human language contains words which are relatively resistant to borrowing and relatively stable over time due to the meaning they express: these words constitute the basic vocabulary of languages.

2. Shared retentions in the basic vocabulary of different languages reflect their degree of genetic relationship.

### 1.2 The Lexicostatistical Working Procedure

In contrast to Dyen *et al.* (1992:95-98), we distinguish five steps for the lexicostatistical working procedure. Due to the fact that in many recent and old applications of lexicostatistics, the actual lists of basic

*This is an updated handout for a presentation given at the Arbeitstagung der Indogermanischen Gesellschaft 2009: Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik. Würzburg. 24.-26.September 2009.

vocabulary items were not solely based on the original meaning lists proposed by Morris Swadesh (cf. Swadesh 1952, 1955), the selection (or compilation) of an appropriate list of basic concepts should be included in a description of the lexicostatistical working procedure.

1. **Swadesh-List Compilation**: Compile a list of basic vocabulary items (a Swadesh list)
2. **Swadesh-List Translation**: Translate the items into the languages that shall be investigated
3. **Cognate Judments**: Search the language entries for cognates
4. **Cognate Coding**: Convert the cognate information into a numerical format
5. **Computation**: Compute a graphical representation out of the numerical data

Up to today, dozens of different Swadesh-Lists have been compiled for various purposes. Swadesh-Lists of all kinds are used as heuristical tools for the detection of deep genetic relationships among languages (c.f. e.g. Dolgopolsky 1986), as basic values for traditional lexicostatistical and glottochronological studies (cf. e.g. Gray & Atkinson 2003), or as litmus test for dubious cases of language relationship which might be due to inheritance or borrowing (cf. e.g. McMahon & McMahon 2005, Chen 1996, Wang 2006). The list of different Swadesh-Lists in Table 1 which have been proposed so far is not exhaustive. Our database lists close to 50 different Swadesh-Lists (and we are sure that we have not yet been able to find all of them).

| **Swadesh-List** | **Source** | **Description** |
|---|---|---|
| Matisoff-200 | Matisoff 1978 | Swadesh-List for lexicostatistical applications on Sino-Tibetan Languages |
| Blust-210 | Greenhill *et al.* 2008 | Swadesh-List for Austronesian languages |
| Swadesh-200 | Swadesh 1952 | The first broadly recognized Swadesh-List |
| Swadesh-100 | Swadesh 1955 | The revision of Swadesh-200 |
| Starostin-110 | Starostin 1999 | The traditional list used for the more than 400 languages in the Tower of Babel project, based an a merger of Jachontov-100 (unpublished, cf. Starostin 1999) and Swadesh-100 |

Table 1: Some Examples for different Swadesh-Lists

## 1.3 Main Critics Regarding Lexicostatistics

Soon after Morris Swadesh established lexicostatistics as a new method in historical linguistics, the method was critized in many publications for all its obvious shortcomings. In the recent applications of lexicostatistics, most of these critics are explicitly mentioned and commented by Swadesh's new followers (cf. Table 2). In this context, it is interesting to note, that - to our knowledge - the last point of criticism has not yet been explicitly addressed in the recent lexicostatistical literature. This coincides well with a general tendency in studies concerning lexicostatistics (even the critical ones) to ignore the data

| **Critic** | **Author** | **Reply** | **Author** |
|---|---|---|---|
| Distances do not tell us anything about language history. | Blust 2000 | Our methods are character-based | Atkinson & Gray 2006 |
| Borrowing will make the results unreliable | Bergsland & Vogt 1962 | Not within basic vocabulary | Atkinson & Gray 2006 |
| Basic vocabulary is not resistant to borrowing | Sagart & Lee 2008 | In most cases it still is | Starostin 1999 |
| The method and its data basis is subjective and inconsistent | Hoijer 1956, Rea 1973 | NO REPLY SO FAR | |

Table 2: Some Critics Regarding Lexicostatistics

basis almost completely, safely assuming that possible errors in translation and coding won't turn out to be statistically significant. A popular example is Tischler & Ganter's (1997) review of Dyen *et al.* (1997), where the authors comment the data basis as follows:

> Besagte Zahlenwerte (Prozentsätze der Übereinstimmungen im Grundwortschatz) wurden unter Verwendung der bekannten, 200 Begriffe enthaltenden Swadesh'schen Wortliste, ermittelt. Ihre Richtigkeit ist zwar nicht überprüfbar, da die Werte sich jedoch im Rahmen der von anderen Untersuchungen bekannten und durch eigene Versuche ermittelten Daten bewegen, seien sie hier nicht weiter angezweifelt. (Tischler & Ganter 1997:44)

It is surprising that a scholar like Tischler, one of the few experts in historical linguistics who also are experienced in lexicostatistics and glottochronology, reveals such a trust in lexicostatistical datasets. From his study on the validity of lexicostatistics and glottochronology from the 1970s we assume that he was well aware of the fact that in the literature there are numerous examples of differences in the results of lexicostatistical analyses carried out by different researches on the same set of languages. Tischler himself mentioned some of these cases in his thesis from 1973 (Tischler 1973, 119f).

## 2 Data Problems

Let us start with some general considerations regarding possible shortcomings of lexicostatistical datasets. Due to the fact that parts of the lexicostatistical working procedure are based on individual decisions which might be prone to subjectivism, we expect to find the greatest problems within step 2 (item translation) and step 3 (cognate judgments) of the lexicostatistical working procedure. We can distinguish two kinds of possible errors in these two steps of the lexicostatistical working procedure: Methodological errors, i.e. errors provoked by shortcomings of the method, and individual errors, i.e. errors provoked by shortcomings of individual scholars applying the method.

Regarding step two of the lexicostatistical working procedure, we identify the following methodological and individual sources of errors:

- **Methodological Errors**:
  - conceptual fuzziness
  - synonymous differentiation in the target languages
  - linguistic diversity
- **Individual Errors**:
  - lack of competence in the target language
  - use of low-quality references

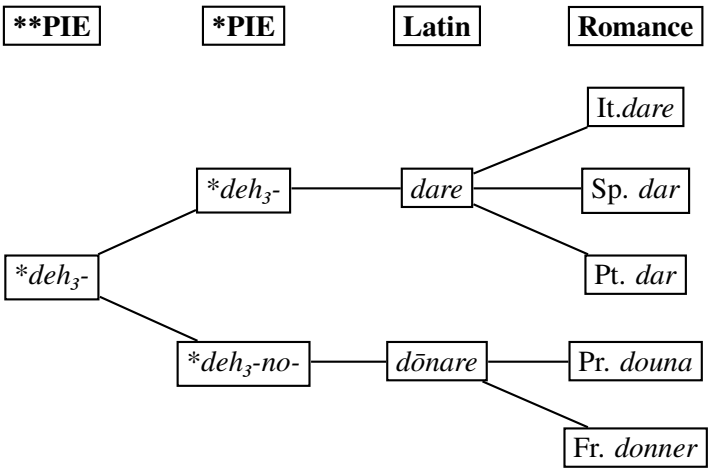


Figure 1: The Problem of Reconstruction Depth

Regarding possible problems of cognate judgments, a specific problem in lexicostatistics is that the questions of reconstruction depth has never been solved sufficiently. What should count as a cognate: Language entries which can be matched completely, i.e. the few examples which we have in historical linguistics, where sound changes took place without the slightest exception? Or should we base the cognate judgments on root-identity, as it is the usual practice in many lexicostatistical applications? But what does the fact, that items do *not* match, tell us then? The fundamental idea of lexicostatistics is that replacements of word forms in certain meaning slots of the basic part of the lexicon constitute a regular process. If we consider the forms for the basic item "give" in Figure 1, it is obvious that we are dealing with a real replacement of the form Lt. *dare* in Provencal and French, since the etymological connection between Lt. *dōnare* and *dare* was surely *not* transparent for the Romans. From a root-perspective, however, we have to count all forms as cognates: they go all back to the PIE root *$deh_3$ "give" (cf. Meiser 1999).

# 3 Comparison of Lexicostatistical Datasets

## 3.1 Our Data

To check to which degree the problems of methodological and individual errors in lexicostatistical datasets may influence the results of computer-analyses, we have compiled a comparative dataset of two large lexicostatistical databases for Indo-European, namely the Dyen database (cf. Dyen *et al.* 1997) and the

Tower of Babel (cf. Starostin 2008) database. In order to have two independent test lists provided by different scholars which are maximally comparable we extracted a set of 46 languages and 103 basic vocabulary items which occur in both datasets. The cognate judgments for the Dyen database are based on the application of Russel Gray (cf. Gray & Atkinson 2003), which we further compared with the cognate judgments displayed in the original dataset.

In order to make the datasets comparable, we applied the following steps:

- **Intersection of both datasets**: We chose only those languages and entries which would overlap in both datasets, this was the only reason for the selection of items and languages.
- **Making the coding similar**: Both loans and gaps were coded by assigning negative numbers to the words (this is the usual practice in the STARLING-software package, cf. Starostin 1993, which we were using for a part of our calculations).
- **Excluding singletons**: All singletons were excluded from the analysis, i.e. all words which were not cognate to any other word in the text (this was necessitated by the coding of the Dyen database which follows exactly this procedure, Tower of Babel differs in several respects from Dyen, so we changed the coding of Tower of Babel according to the Dyen standards, since this was the only way to make the data comparable without applying our own decisions)
- **Restricting cognate judgments to item identity**: Tower of Babel assigns the same number to all etymologically related words, so English "what" and "who" will be given the same number. Since the Dyen database was not coded in this way, we replaced all numbers which would show up in different rows of items by new numbers

| **Author** | **Dyen *et al.* 1997** | **Tower of Babel (no date)** | **Intersection** |
|---|---|---|---|
| Language family | Indo-European | Indo-European | Indo-European |
| Number of lang. | 95 | 98 | 46 |
| Number of items | 200 | 110 | 103 |

Table 3: The Structure of the Two Datasets

## 3.2 Coding Trouble

### The Dyen-Database

The trouble with the encoding in the Dyen database is that the problem of multiple language entries was not solved properly. Instead of allowing to list multiple entries separately, Dyen *et al.* (1997) applied a strange method of assigning relation codes (codes preceded by 'c' in Figure 2) to pseudo-cognatesets (all language entries listed under a specific cognate header, preceded by 'b' in Figure 2), which in turn lead to non-transitive cognate judgments, as illustrated in Figure 3: The cognate sets going back to two distinct Latin forms (*avis* 'bird' and *passer* 'sparrow') are interlinked by the 'c'-lines, only because there are two entries in Spanish, each corresponding to one of the two Latin roots. These cognate judgments are very hard to check on their correctness. In order to compile the data for the biological software packages, one has to untangle the 'networks of cognacy' proposed by the authors, which is a task that, unfortunately, cannot be done in a consistent way. The confusing network of all inter-cognate relationships in the Dyen-Database gives a rough approximation of the complexity of the data (Figure 3).

| Cogn.-ID | Cognate-Relation | Mean.-ID | Lang.-ID | Language | Language-Entry |
|---|---|---|---|---|---|
| b 200 | | | | | |
| | c 200 2 201 | | | | |
| | | 012 | 10 | Italian | **UCCELLO** |
| | | 012 | 15 | French Creole C | ZIBYE,**ZWEZO** |
| | | 012 | 23 | Catalan | **AUCELL**,MOIXO |
| | | 012 | 12 | Provencal | **AUCEU** |
| | | 012 | 13 | French | **OISEAO** |
| | | 012 | 21 | Portuguese ST | **AVE** |
| b 201 | | | | | |
| | c 200 2 201 | | | | |
| | c 201 2 202 | | | | |
| | | 012 | 20 | Spanish | **AVE**,PAJARO |
| b 202 | | | | | |
| | c 201 2 202 | | | | |
| | | 012 | 08 | Rumanian List | PASARE |

Figure 2: The Coding of the Dyen-Database

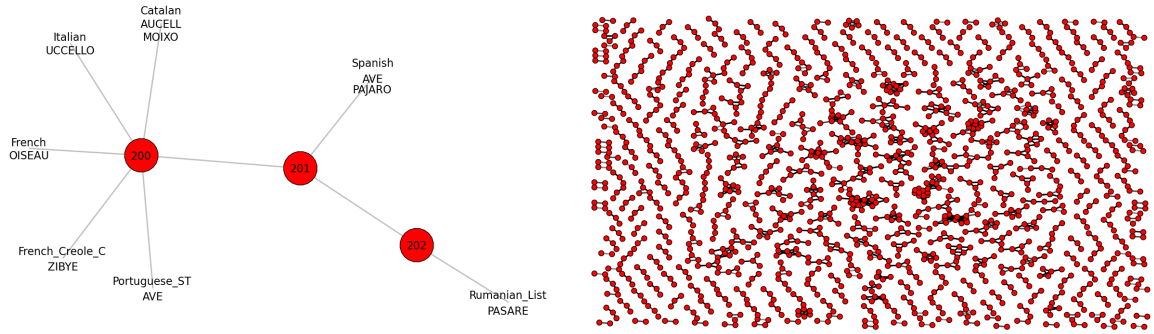


Figure 3: Inter-Cognate-Relations in the Dyen-Database (Example BIRD and Full Set of Cognates)

### The Tower-of-Babel-Database

Tower of Babel created a special way of encoding lexicostatistical word-lists which is implemented in the STARLING software package (cf. Starostin 1993). The idea is to simply assign the same number to related entries and to link these entries with proto-forms (which are in fact whole etymological dictionaries). This system is exemplary, both in transparency of cognate judgments and applicability.

## 3.3 Detailed Comparison of the Databases

Table 4 shows a detailed comparison of the entries for BIRD in ToB and the Dyen-Database. In this case, there is only a difference in one item, namely the additional entry for BIRD in Portuguese *passaro*. These

| **Language** | **Language-Entry** | **Cognate-ID** | **Language-Entry** | **Cognate-ID** |
|---|---|---|---|---|
| Latin | ave | 1140 | | |
| Italian | uccello | 1140 | | |
| French | oiseau | 1140 | | |
| Portuguese | ave | 1140 | passaro | 1985 |
| Spanish | ave | 1140 | pajaro | 1985 |
| Provencal | aucel | 1140 | | |
| Romanian | pasăre | 1985 | | |

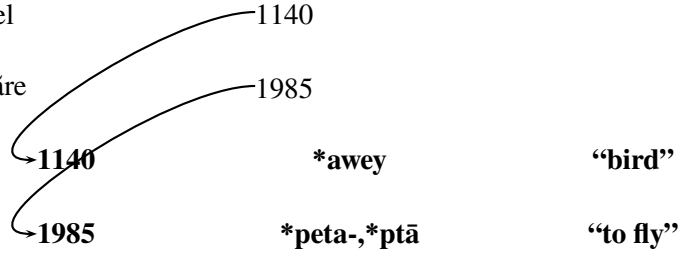


Figure 4: The Coding of Tower of Babel

apparently minor differences, however, sum up to about 10 percent in the whole Romance partition of both databases. This clearly shows that item translation is a huge problem of lexicostatistics. If the datasets which different scholars use in order to draw their conclusions differ to such a great extent, it is almost impossible to compare their results and map them to 'real' historical scenarios of language development.

| **BIRD** | **Dyen** | **ToB** | | **G&L** | |
|---|---|---|---|---|---|
| ita. | UCCELLO | uccello | | uccello | passero |
| fre. | OISEAU | oiseau | | oiseau | passereau |
| port. | AVE | ave | passaro | ave | pássaro |
| spa. | AVE, PAJARO | ave | pajaro | ave | pájaro |
| prov. | AUCEU | aucel | | aucel | paser |
| rom. | PASARE | | pasăre | | pasăre |

Table 4: Comparison of BIRD in ToB and Dyen

## 3.4 Undetected Borrowings

While differences in item translation can surely be considered as an inherent problem of lexicostatistical methodology and thus belonging to our category of "methodological errors", the many cases of undetected borrowings which we could identify in both datasets (although the Dyen-Database performed worse), clearly belong to the latter category of individual errors. Table 5 is a non-exhaustive list of some of the most typical cases of undetected borrowings within the Romance partition of both datasets.

| Author | Item | Donor | Quelle | rom. | it. | pr. | fr. | sp. | pt. |
|---|---|---|---|---|---|---|---|---|---|
| **Dyen** | KILL | fr. | tuer | | | tua | | | |
| | ROAD | gr. | drómos | drum | | | | | |
| | ROAD | ir. | strada | stradă | | | | | |
| | ROAD | fr. | rue | | | | | | rua |
| | SKIN | lt. | cutis | | | | | cutis | |
| | WALK | frk. | marka | | | marcha | marcher | | |
| | WOMAN | gr. | familia | femeie | | | | | |
| **ToB** | TAIL | lt. | cauda | | | | | | cauda |
| | THIN | fr. | mince | | | mince | | | |
| | WARM | lt. | calidus | | calido | | | | |
| | WOMAN | gr. | familia | femeie | | | | | |
| | KILL | fr. | tuer | | | tuar | | | |

Table 5: Undetected Borrowings in the Dyen-Database and Tower of Babel

## 3.5 Tree Topologies of the Whole Datasets

How do the differences we identified in the two datasets surface when applying step 5 of the lexicostatistical working procedure and computing family trees out of the coded data? In order to test this we applied several methods of tree conversion, using distance- and character-based approaches. In order to have a first rough approximation of differences, we measured the split-differences between the trees, using the TOPD-software (cf.Puigbò *et al.* 2007). These comparisons reveal, that all computed tree topologies differ by 30 - 40 % regarding their splits. The results for the Bayesian analyses [1], which performed best, showing split differences of only about 30 %, are given in Figures 5 and 6. A closer comparison of these two figures clearly shows, that the differences between the two trees are so great that they cannot be simply ignored. These differences occur in all parts of the trees, showing conflicts in higher phylogenies and in the subgrouping of closer related languages. Note that these differences are only due to differences in cognate judgments and item translations. Both datasets contain the same number of items and the same number of languages, so actually - assuming that lexicostatistics is a valid method - they should show no differences at all.

[1] Analysis was made using MrBayes (cf. Ronquist & Huelsenbeck 2003), noabsencesites for the rates, gamma for the encoding, and Albanian as an outgroup. 1.5 million trees of both datasets were created (by this time, both datasets had reached convergence), of which we sampled 1000 for the consensus trees (burn in was 250)

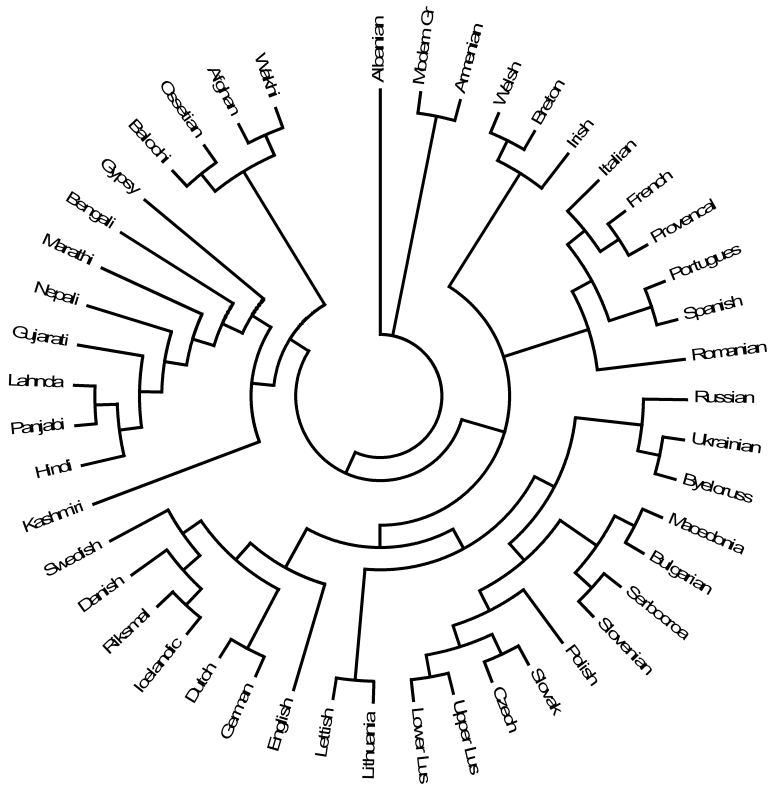


Figure 5: The Bayesian Analysis of the Tower-of-Babel-Dataset

| **Method** | **Split-Difference (%)** |
|---|---|
| Traditional | 39.53 |
| Matching | 32.56 |
| Jaccard | 37.21 |
| Correlation | 44.19 |
| Cosine | 41.86 |
| MrBayes | 30.23 |

Table 6: Split-Differences between the Dyen and ToB

## 4 Back to the Roots

What is left to say? Does lexicostatistics have a future, or was Rea (1973:361) right in his pessimistic resumee:

> If, as Lees and Chrétien feel, the mathematics are inadequate; if as Hall, Bergsland and Vogt, Arndt, O'Neill, Coseriu, Fodor, I and others have found, the results of the method do

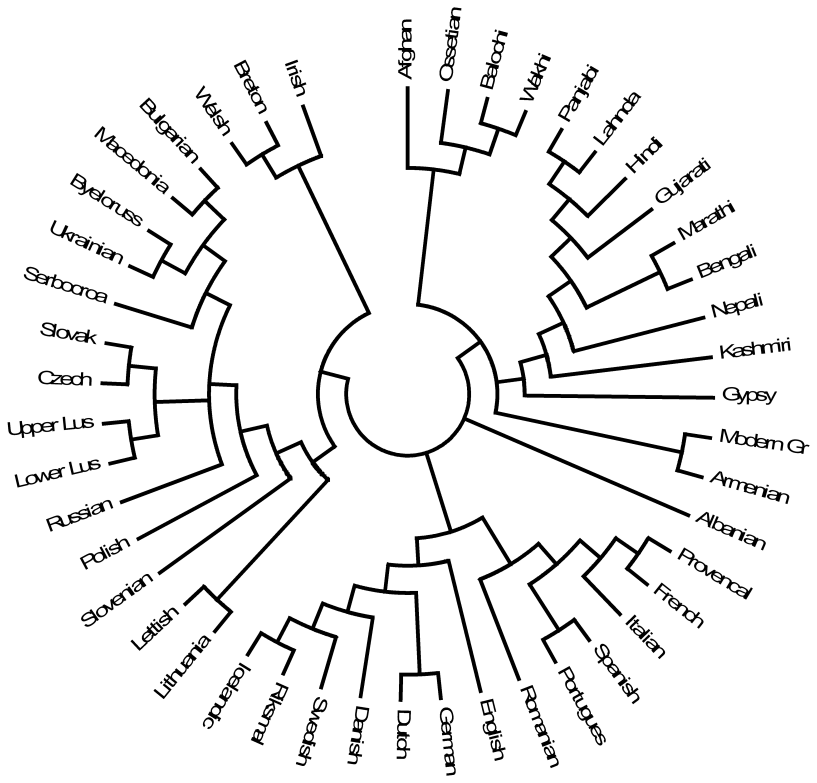


Figure 6: The Bayesian Analysis of the Dyen-Dataset

> not correspond to known facts, if now, the Romance wordlists and scorings that formed the basis of the method are in fact full of indeterminencies, inconsistencies and errors, what then remains?

We think that lexicostatistics in its current form does not have a future, but we do not think that because of this failure of one particular method, all quantitative approaches should be given up all at once. We especially hope that root-based approaches which are closer to traditional methodology of historical linguistics (cf. e.g. Starostin 2000, Holm 2007, Ellegård 1959) will produce datasets which are less prone to subjective judgments and individual errors. Datasets encoded in this way can then further used for phylogenetic calculations in the EvoClass research project, and we hope that they will provide a more objective basis for stochastic calculations on linguistic datasets and may reveal interesting aspects of language history.

## References

Arapov, M. V., & M. M. Cherc. 1983. *Mathematische Methoden in der historischen Linguistik*. Brockmeyer. Translator: Köhler, R. und Schmidt, P.

Atkinson, Quentin D., & Russell D. Gray. 2006. How old is the indo-european language family? illumination or more moths to the flame? In *Phylogenetic methods and the prehistory of languages*, ed. by Peter Forster & Colin Renfrew, McDonald Institute monographs, 91–109, Cambridge UK , Oxford UK , Oakville CT USA ,. McDonald Institute for Archaeological Research; Distributed by Orbow Books.

Bergsland, Knut, & Hans Vogt. 1962. On the validity of glottochronology. *Current Anthropology* 3.115–153.

Blust, Robert. 2000. Why lexicostatistics doesn't work. the 'universal constant' hypothesis and the Austronesian languages. In *Time depth in historical linguistics*, ed. by C. Renfrew, A. McMahon, & L. Trask, p. 311–331. Cambridge: The McDonald Institute for Archaeological Research.

Chen, Baoya. 1996. *Lun yuyan jiechu yu yuyan lianmeng (Language contact and language unions)*. Beijing: Yuwen Chubanshe.

Dolgopolsky, A. B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of northern eurasia. In *Typology Relationship and Time*, ed. by T. L. Shevoroshkin, Vitaly V.; Markey, Notes on Linguistics, 27–50. Karoma Publisher, Inc. Originally published in Russian as "Gipoteza drevnejščego rodstva jazykov Severnoj Evrazii (problemy fonetičeskich sootvetstvij)" in 1964.

Dyen, Isidore, Joseph B. Kruskal, & Paul Black. 1992. An indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82.iii–132.

Dyen, Isidore, Joseph B. Kruskal, & Paul Black, 1997. Comparative indo-european database: File ie-data1.

Ellegård, A. 1959. Statistical measurement of linguistic relationship. *Language* 35.131–156.

Gray, Russell D., & Quentin D. Atkinson. 2003. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature* 426.435–439.

Greenhill, Simon J., Robert Blust, & Russell D. Gray. 2008. The austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4.271–283.

Gudschinsky, Sarah C. 1956[1964[. The abc's of lexicostatistics (glottochronoloy). In *Language in culture and society: A reader in linguistics and anthropology*, ed. by Dell H. Hymes, A Harper international edition. New York: Harper and Row, reprint edition. (Originally published in Word 12: 175-210).

Hoijer, Harry. 1956. Lexicostatistics: A critique. *Language* 32.49–60.

Holm, Hans J. 2007. The new arboretum of indo-european "trees": Can new algorithms reveal the phylogeny and even prehistory of indo-european? *Journal of Quantitative Linguistics* 14.167–214.

Matisoff, James A. 1978. *Variational semantics in Tibeto-Burman. The 'organic' approach to linguistic comparison*. Institute for the Study of Human Issues.

McMahon, April, & Robert McMahon. 2005. *Language classification by numbers*. Oxford: Oxford University Press.

Meiser, Gerhard. 1999. *Historische Laut- und Formenlehre der lateinischen Sprache*. Darmstadt: WBG (Wissenschaftliche Buchgesellschaft).

Puigbò, Pere, Santiago Garcia-Vallvé, & James O. McInemey. 2007. Topd/fmts: a new software to compare phylogenetic trees. *Bioinformatics* 23.1556–1558.

Rea, John A. 1973. The romance data of pilot studies for glottochronology. In *Diachronic, areal and typological linguistics*, ed. by Henry Max Hoenigswald & Robert H. Langacre, volume 11 of *Current Trends in Linguistics*, 355–367. The Hague; Paris: Mouton.

Ronquist, Frederik, & J. P. Huelsenbeck. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19.1572–1574.

Sagart, Laurent, & Yeon-Ju Lee. 2008. No limits to borrowing: The case of bai and chinese. *Diachronica* 25.357–385.

Sankoff, David, 1969. *Historical linguistics as stochastic process*. Montreal: McGill University dissertation. (A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy).

Starostin, George, 2008. Tower of babel. an etymological database project. http://starling.rinet.ru/main.html.

Starostin, Sergej Anatol'evic. 1999. Methodology of long-range comparison. In *Historical linguistics & lexicostatistics*, ed. by Vitaly Shevoroshkin & Paul J. Sidwell, volume 3 of *AHL Studies in the science & history of language*, 61–66. Melbourne: Assoc. for the History of Language.

Starostin, Sergej Anatol'evic. 2000. Comparative-historical linguistics and lexicostatistics. In *Time depth in historical linguistics*, ed. by Colin Renfrew, April McMahon, & Larry Trask, Papers in the prehistory of languages, 223–265. Cambridge: The McDonald Institute for Archaeological Research. Translation: Peiros, Ilia. Originally published in 1989.

Starostin, Sergej Anatol'evič. 1993. Rabočaja sreda dlja lingvista (working environment for a linguist). In *Bazy dannyh po istorii Evrazii v srednie veka*, p. 7–23.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to north american indians and eskimos. *Proceedings of the American Philosophical Society* 96.452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21.121–137.

Tischler, Johann. 1973. *Glottochronologie und Lexikostatistik*. Insbrucker Beiträge zur Sprachwissenschaft. Kowatsch.

Tischler, Johann, & B. Ganter. 1997. Review of i. dyen, j. kruskal & p. black: An indoeuropean classification (1992). *Kratylos* 42.43–50.

Wang, Feng. 2006. *Comparison of languages in contact: The distillation method and the case of Bai*. Taipei: Inst. of Linguistics Academia Sinica.