Beautiful Trees on Unstable Ground

Notes on the Data Problem in Lexicostatistics

Hans Geisler and Johann-Mattis List

Introduction

While lexicostatistics and glottochronology had been suffering a lack of prestige 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 these methods.

Lexicostatistics

Basic Assumptions of Lexicostatistics

Various authors have tried to summarize the basic assumptions of lexicostatistics in a consistent manner. Yet when turning to the more popular accounts on the basic assumptions of the method which are, e. g., given in the work of Arapov & Herz (1983: 17-20), Gudschinsky (1956) and Sankoff (1969: 2f), one realizes that these accounts all differ to a certain degree. We propose that the core of lexicostatistical theory can be summarized 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*, i.e. they are representative for the reconstruction of language phylogenies.

These two basic assumptions introduce the two main ideas of lexicostatistics as they were first proposed by Morris Swadesh in his early papers at the begin of the 1950ies (cf. Swadesh 1950, 1952 & 1955), namely the idea of *basic vocabulary* as a specific set of concepts which are expressed in all languages and the idea that *shared cognates* within the realm of basic vocabulary reflect the *degree of genetic closeness* among languages. In our opinion, all methods for phylogenetic reconstruction which are based on these two basic assumptions can be classified as lexicostatistical approaches.

The Lexicostatistical Working Procedure

In contrast do Dyen *et al.* (1992: 95-98), who explicitly divide the lexicostatistical working procedure into four steps, we distinguish five steps. This is due to the fact that in many recent and old applications of lexicostatistics, the actual lists of basic vocabulary items were not solely based on one of the two original meaning lists proposed by Swadesh (1952 & 1955). Therefore, the selection or compilation of an appropriate list of basic concepts should be included in a description of the lexicostatistical working procedure. Thus, the lexicostatistical working procedure can be characterized by the following five steps:

Step 1: *Compilation*: Compile a list of basic vocabulary items (a Swadesh-list).
Step 2: *Translation*: Translate the items into the languages that shall be investigated.[1]
Step 3: *Cognate Judgments*: Search the language entries for cognates.
Step 4: *Coding*: Convert the cognate information into a numerical format.
Step 5: *Computation*: Perform a computational analysis (cluster analysis, tree calculation) of the numerical data, which allows to make conclusions regarding the phylogeny of the languages under investigation.

Main Critics Regarding Lexicostatistics

Soon after Morris Swadesh established lexicostatistics as a new method in historical linguistics, the method was criticized in many publications for all its obvious shortcomings. Table 1 lists a few of the most crucial points which have been discussed so far, along with a reply by modern practitioners of lexicostatistics.

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

Table 1: Some critics regarding lexicostatistics

In recent applications of lexicostatistics, most of these critics are explicitly mentioned and commented by Swadesh's new followers, as Table 1 shows. Yet 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 underestimate the importance of the

[1] We use the term "translation" in this context, since it is traditionally used in the lexicostatistical literature for the process of finding a word in a certain language which expresses a given basic concept.

data basis for lexicostatistical analyses, safely assuming that possible errors in translation and coding won't turn out to be statistically significant.

Problems of Data Handling

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* (*Translation*) and *Step 3* (*Cognate Judgments*) of the lexicostatistical working procedure. We can distinguish two kinds of possible errors in these two steps: 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 2* of the working procedure, we identify the following methodological sources of errors:

1: *Concept Fuzziness*: The basic concept is defined in a way that makes it difficult to find a unique, best match for the translation into the target language. Cf. e.g. the basic concept KNOW for which it is very difficult to decide which of the possible German equivalents *kennen* "know something", *wissen* "know facts" would match it best.

2: *Synonymous Differentiation*: The target language offers more than one translation for the basic concept due to language specific differentiation. Cf. e.g. the two possible translations for the basic concept BIRD in Spanish, where *pájaro* refers to small birds, while *ave* refers to big birds.

3: *Linguistic Diversity*: The target language offers different translations for the basic concept, due to dialectal or sociolinguistic variation. Cf. e.g. the two possible translations for the basic concept KILL in German, where we have *umbringen* as a colloquial and *töten* as a literal variant.

As individual sources of errors, we identify the following two:

1: *Lack of Competence*: If the researcher doesn't have a sufficient knowledge of the target language, which is necessarily the case when – as Dyen et al. (1997) did – a handful of researchers tries to analyze a set of 95 languages, many of which are only sparsely documented, errors in the coding will be unavoidable.

2: *Use of Low-Quality References*: Errors will likewise increase, if the references which are taken into account when translating the basic concepts into the target languages are of low quality or out-dated.

Regarding possible problems of cognate judgments (*Step 3*), a specific problem in lexicostatistics is that the question of reconstruction depth has never been solved sufficiently. What should count as cognate: Language entries which can be matched completely, i.e. the few examples which we have in historical linguistics, where regular 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 Latin *dare* "give" in Provencal and French, since the etymological connection between Latin *dōnare* "give as a present" and *dare* "give" was surely not transparent for the Romans. From a root-perspective, however, we have to count all forms as cognates: they all go back to the PIE root **deh*$_3$- "give" (cf. Meiser 1999).

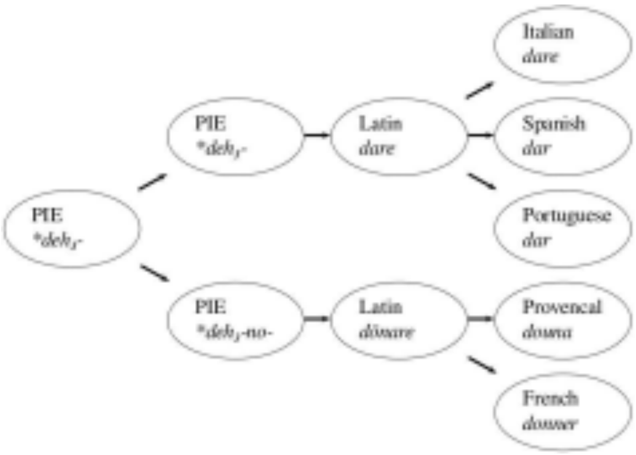


Figure 1: The problem of reconstruction depth

To sum up: The lexicostatistical working procedure and its above-mentioned shortcomings lead to datasets where we are dealing with an arbitrary selection of language variants with arbitrary assignments of cognacy.

Comparison of Datasets for Lexicostatistics

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 compared of two large lexicostatistical databases for Indo-European, namely the *Dyen* database (cf. Dyen *et al.* 1997) and the *Tower of Babel* database (cf. Starostin 2008). 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 Gray & Atkinson (2003) which we further compared with the cognate judgments displayed in the original dataset. The cognate judgments for the *Tower of Babel* dataset were provided by the *Tower of Babel* team.

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

1: *Intersection of both datasets:* We chose only those languages and basic vocabulary items which would overlap in both datasets. This was the primary reason for the selection of basic vocabulary items and languages.
2: *Making the coding similar:* Both loans and gaps were coded by assigning negative numbers to the respective translated entries (this is the usual practice in the STARLING software package, cf. Starostin 1993), which we used for part of our calculations.
3: *Excluding "singletons":* All singletons, i.e. all translated entries which are not cognate with any other entries, were excluded from the analysis.
4: *Restricting cognate judgments to item identity:* Tower of Babel assigns the same cognate ID to all etymologically related words, such that e.g. English *what* and *who* are given the same ID. Since the *Dyen* database was not coded in this way, we changed the coding of the *Tower of Babel* database.

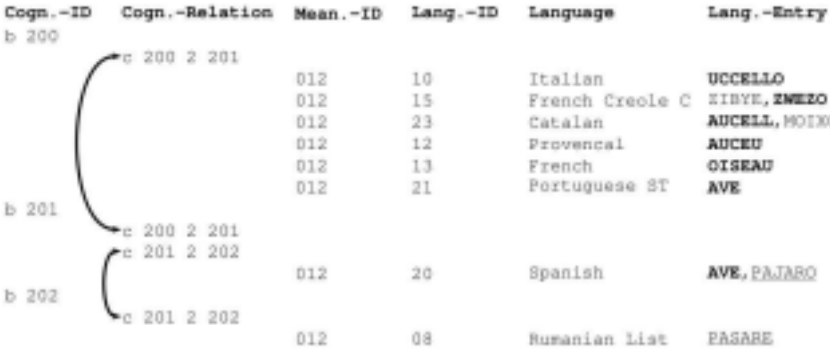
Table 2 gives an overview over the way we coded both databases in order to make them comparable.

Database	**Dyen et al. 1997**	**Tower of Babel**	**Intersection**
Language family	Indo-European	Indo-European	Indo-European
No. of languages	95	98	46
No. of items	200	110	103

Table 2: The structure of the two datasets

Coding Problems in the Dyen Database

The trouble with the encoding in the *Dyen* database is that the problem of multiple language entries was not solved properly. Figure 2 gives an example on the way the data is coded in this database.



Cogn.-ID	Cogn.-Relation	Mean.-ID	Lang.-ID	Language	Lang.-Entry
b 200					
	c 200 2 201				
		012	10	Italian	**UCCELLO**
		012	15	French Creole C	XIBYE, **ZWEZO**
		012	23	Catalan	**AUCELL**, MOIXO
		012	12	Provencal	**AUCEU**
		012	13	French	**OISEAU**
		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

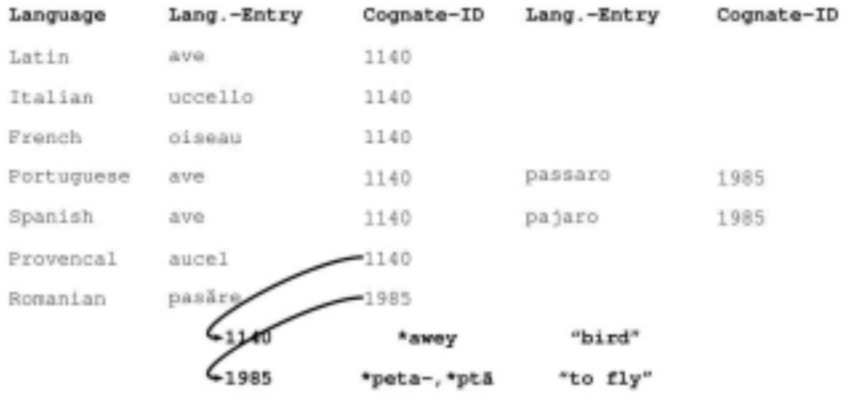
Figure 2: Coding of the *Dyen* dataset

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 cognate sets (all language entries listed under a specific cognate header, preceded by "b" in

Figure 2), which in turn lead to non-transitive cognate judgments: The cognate sets going back to two distinct Latin words denoting two distinct concepts (*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 result is a confusing network of inter-cognate relationships.

The Coding of the Tower of Babel Database

The *Tower of Babel* project 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, i.e. cognate, entries and to link these entries with proto-forms, which makes these databases to complete etymological dictionaries.



Language	Lang.-Entry	Cognate-ID	Lang.-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		
	1140	*awey	"bird"	
	1985	*peta-, *ptă	"to fly"	

Figure 3: Coding of the *Tower of Babel* dataset

This system, which is reflected in Figure 3, is exemplary, both in transparency of cognate judgments and applicability.

Detailed Comparison of the Datasets

In order to get a first impression regarding the differences in the datasets which can be explained by the sources of errors we identified, we carried out a closer examination of the Romance partition of both datasets. As an example, Table 3 gives a detailed comparison of the entries for BIRD in *Tower of Babel* and the *Dyen* database. In this case, there is only a difference in one item, namely the additional entry for BIRD in Portuguese in the *Tower of Babel* dataset.

BIRD	***Dyen***	***Tower of Babel***	
Italian	UCCELLO	uccello	
French	OISEAU	oiseau	
Portuguese	AVE	ave	passaro
Spanish	AVE,PAJARO	ave	pajaro
Provencal	AUCEU	aucel	
Romanian	PASARE		pasăre

Table 3: Comparison of BIRD in *Dyen* and *Tower of Babel*

These 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.

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 4 gives a non-exhaustive list of some of the most typical cases of undetected borrowings within the Romance partition of both datasets.[2]

Author	**Item**	**Donor**	**Source**	**Recipient Languages**				
		Language	**Lang.**	**Rom.**	**Italian**	**Provencal**	**French**	**Spanish**
Dyen	KILL	French	*tuer*			tua		
	ROAD	Greek	*drómos*	drum				
	SKIN	Latin	*cutis*					cutis
	WALK	Old Franc.	*marka*			marcha	marcher	
	WOMAN	Greek	*familia*	femeie				
ToB	THIN	French	*mince*			mince		
	WARM	Latin	*calidus*		calido			
	WOMAN	Greek	*familia*	femeie				
	KILL	French	*tuer*			tuar		

Table 4: Undetected borrowings in *Dyen* and *Tower of Babel*

Comparing the Computed Tree Topologies of the 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

[2] Rumanian *femeie* is misjudged in both datasets for being cognate to French *femme*, Sardinian *femmina*, etc. Only the ladder go back to Latin *femina*, whereas the Rumanian word is clearly related to Latin *familia* which was first borrowed into Turkish and changed meaning from "family" to "woman (in a harem)". The word migrated with his new meaning from Turkish to Greek and then to Rumanian.

et al. 2007). Table 5 lists the results of these tests for different methods we applied for the data analysis.

Method	Split-Difference (%)
Uncorrected distances (Neighbor-Joining)	39.53
Cosine distances (Neighbor-Joining)	32.56
Matching distance (Neighbor-Joining)	41.86
MrBayes (Bayesian approach)	30.23

Table 5: Split differences between *Dyen* and *Tower of Babel*

These comparisons reveal that all computed tree topologies differ by 30 – 40% regarding their splits. Note that – for the Romance partition of both datasets, where our analysis revealed about 10 percent differences in item translation and cognate judgments – the tree topologies in both analyses are the same. Thus, the differences which we identified even do not show up in the tree topologies, suggesting that the differences in the rest of the datasets are even greater than in the Romance part.

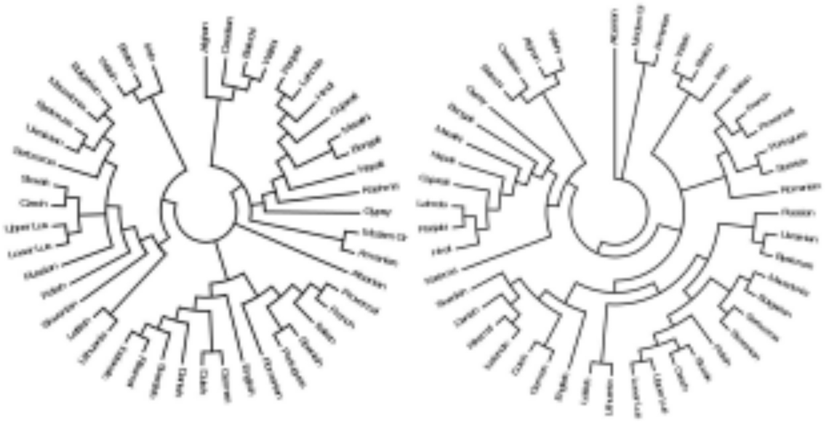


Figure 4: Bayesian analysis for *Dyen* (left) and *Tower of Babel* (right)

The results for the Bayesian analyses[3], which performed best, showing split differences of only about 30%, are given in Figure 4. 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 – there should be no differences at all.

[3] The analysis was made using the MrBayes software package (cf. Ronquist & Huelsenbeck 2003) with Albanian as 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.

Conclusion: Back to the Roots?

What is left to say? Our analysis clearly shows that differences in item translation and cognate judgments have a great impact on the topology of the trees calculated from lexicostatistical datasets. The impression that – due to the large amount of data employed – these differences would not show up in the calculations has proven to be problematic. This shows clearly that the main problem of lexicostatistics lies not in its basic assumptions, which most scholars still see as the most problematic aspect of the method, but in its working procedure which is prone to subjectivism and errors. Given these facts, we may ask whether lexicostatistics has a future after all, or whether John Rea was right in his pessimistic résumé, stated about forty years earlier:

> 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 not correspond to known facts; if now, the Romance wordlists and scorings that formed the basis of the method are in fact full of indeterminacies, inconsistencies and errors, what then remains? (Rea 1973: 361)

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 to genetic language classification should be given up at once. We especially hope that root-based approaches which are closer to the traditional methodology of historical linguistics (cf. e.g. Starostin 2000, Holm 2002) 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, and we hope that they will provide a more objective basis for stochastic calculations on linguistic datasets and may reveal interesting aspects and new insights into the complexity of language history.

Bibliography

M. M. Arapov and M. M. Cherc, Mathematische Methoden in der historischen Linguistik, Bochum 1983.

Quentin D. Atkinson and Russel D. Gray, How old is the Indo-European language family? Illumination or more moths to the flame? in: Phylogenetic methods and the prehistory of languages, ed. by P. Forster & C. Renfrew, Cambridge 2006: 91-109.

Knut Bergsland and Hans Vogt, On the validity of glottochronology, in: Current Anthropology 3 (1962): 115-153.

Robert Blust, Why lexicostatistics does'nt work. The 'universal constant' hypothesis and the Austronesian languages, in: Time depth in historical linguistics, ed. by C. Renfrew *et al.*, Cambridge 2000: 311–331.

Isidore Dyen *et al.*, An Indoeuropean classification: A lexicostatistical experiment, in: Transaction of the American Philosophical Society 82 (1992): iii–132.

Isidor Dyen *et al.*, Comparative Indoeuropean database: File IE-data1, 1997 (Online available under: http://www.wordgumbo.com/ie/cmp/iedata.txt).

Russel D. Gray and Quentin D. Atkinson, Language tree divergence times support the Anatolian theory of Indo-European origin, in: Nature 426 (2003): 435–439.

Sarah C. Gudschinsky, Three disturbing questions concerning lexicostatistics, in: International Journal of American Linguistics 22 (1956): 212–213.

Harry Hoijer, Lexicostatistics: A critique, in: Language 32 (1956): 49–60.

Hans J. Holm, Genealogy of the main Indo-European branches applying the separation base method, in: Journal of Quantitative Linguistics 7 (2002-2): 73–95.

Pere Puigbò *et al.*, Topd/fmts: A new software to compare phylogenetic trees, in: Bioinformatics 23 (2007): 1556–1558.

John A. Rea, The Romance data of pilot studies for glottochronology, in: Diachronic, areal and typological linguistics, ed. by Henry M. Hoenigswald & Robert H. Langacre, The Hague 1973: 355–367.

Frederik Ronquist and J. P. Huelsenbeck, MrBayes 3: Bayesian phylogenetic inference under mixed models, in: Bioinformatics 19 (2003): 1572–1574.

David Sankoff, Historical linguistics as stochastic process, Montreal 1969.

George Starostin, Tower of Babel: An etymological database project, Moscow 2008 (Online available under: http://starling.rinet.ru).

Sergej A. Starostin, O dokozatel'stve jazykovogo rodstva (On the proof of genetic relationship of languages), in: Tipologija i teorija jazyka (Typology and language theory), ed. by Jekaterina Rachilina, Moskva 1999: 57–69.

Sergej A. Starostin, Comparative-historical linguistics and lexicostatistics, in: Time depth in historical linguistics, ed. by Colin Renfrew *et al.*, Cambridge 2000: 223–265.

Sergej A. Starostin, Rabočaja sreda dlja lingvista (Working environment for a linguist), in: Bazy dannyh po istorii Evrazii v srednie veka (Databases for the history of Eurasia in the Middle Ages), 1993: 7-23.

Morris Swadesh, Salish internal relationships, in: International Journal of American Linguistics 16 (1950): 157–167.

Morris Swadesh, Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos, in: Proceedings of the American Philosophical Society 96 (1952): 452–463.

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

Hans
Geisler
Universitätsstraße 1
40225 Düsseldorf
geisler@phil.uni-duesseldorf.de

Johann-Mattis
List
Universitätsstraße 1
40225 Düsseldorf
listm@phil.uni-duesseldorf.de