Beautiful Trees on Unstable Ground

Notes on the Data Problem in Lexicostatistics

Hans Geisler / Johann-Mattis List (Heinrich Heine University Düsseldorf)[1]

1. Introduction: Lexicostatistics

1.1. Key Assumptions (Swadesh 1950, 1952 & 1955, Lees 1953, Starostin 1999)

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

1.2. The Lexicostatistical Working Procedure (Burlak & Starostin 2005, Dyen 1992)

I. Compile a list of basic vocabulary items (meaning list, Swadesh-list)
II. Translate the items into the languages that shall be investigated
III. Search the language entries for cognates
IV. Convert the cognate information into a numerical format
V. Compute a graphical representation (usually an acyclic, directed graph, i.e. a tree) out of the numerical data

A Short and Non-Exhaustive List of Meaning-Lists[2]:

Matisoff-200	Matisoff 1978 & 2000	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list used for the more than 400 languages in the Tower of Babel project, based an a merger of Jachontof-100 (unpublished, cf. Starostin 1999) and Swadesh-100
Wiktionary-207	Wikipedia's Wiktionary	Simple merger of Swadesh-100 and Swadesh-200, used for the Swadesh-List in Wikipedia's Wiktionary

1.3. Main Critics Regarding Lexicostatistics

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	Lee & Sagart 1999 & 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	

Tischler & Ganter (1997: 44) regarding the data basis of Dyen et al. (1992):

- "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."

Some examples for differences in the data compiled by different scholars:

[1] Contact: geisler@phil-fak.uni-duesseldorf.de, mattis.list@phil-fak.uni-duesseldorf.de

[2] Our project's collection of Swadesh lists currently contains the documentation of 38 Swadesh-Lists (sublists included), and there are many, we have not yet recorded.

- Milke (1962) differs from Bergsland & Vogt (1962) regarding cognate judments and word choice, arriving at different retention rates.
- Swadesh (1962) gets different retention rates form that proposed by Bergsland & Vogt (1962).
- The original test-lists of Swadesh and Lees for the determination of the universal retention rates differed in many points, regarding cognate judgments and item translation (cf. the detailed examples in Rea 1973)

2. Data Problems

2.1. Item Translation (Step 2)

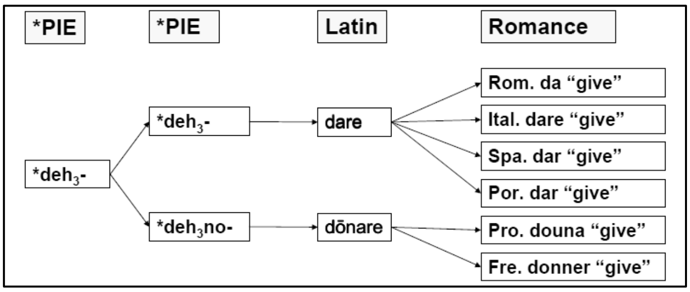
Methodological Errors:

- conceptual fuzziness
- synonymous differentiation in the target languages
- linguistic diversity

Implementation Errors:

- lack of competence in the target language
- use of low-quality references

2.2. Cognate Judments (Step 3, cf. Meiser 1999 for the PIE proto-forms)



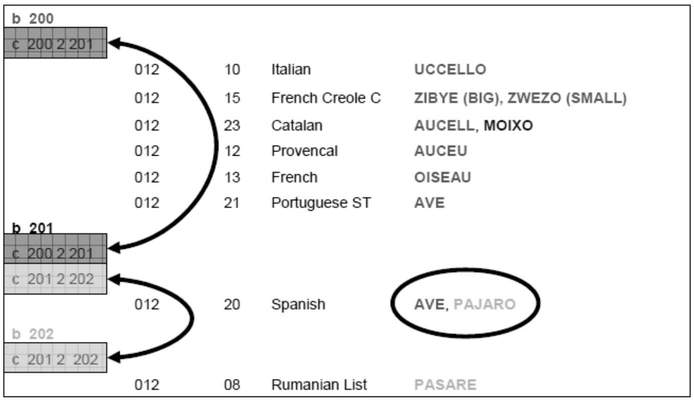
3. Part II: Swadesh-List Comparison

3.1. Comparison of Two Independently Compiled Lexicostatistical Datasets[3]

Author	**Dyen , Kruskal & Black (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

[3] See the Appendix on details of how the datasets were made comparable. For recent publications based on the Dyen-Dataset, cf .e.g. Atkinson et al. (2008), Atkinson et al. (2005), Atkinson & Gray (2006), Searls (2003), Gray & Atkinson (2003), McMahon & McMahon (2005), McMahon & McMahon (2006), Pagel et al. (2007), Rexova et al. (2003), Serva & Petroni (2008).

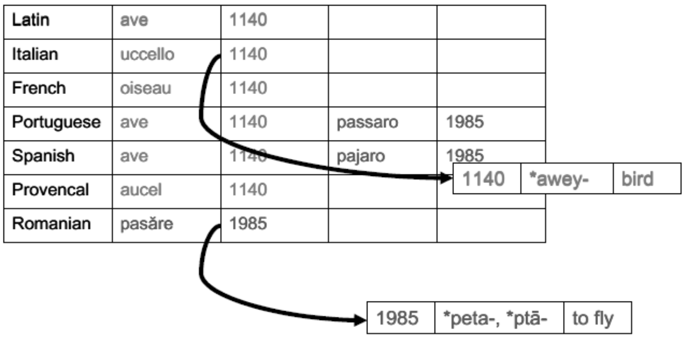
3.2. Dyen et al. (1997): BIRD



```
b 200
c 200 2 201
      012   10   Italian            UCCELLO
      012   15   French Creole C    ZIBYE (BIG), ZWEZO (SMALL)
      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
```

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 to pseudo-cognatesets, which in turn lead to intransitive cognate judgments, which are very hard to check on their correctness.

3.3. Tower of Babel (no date): bird



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-, *ptā-	to fly

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.

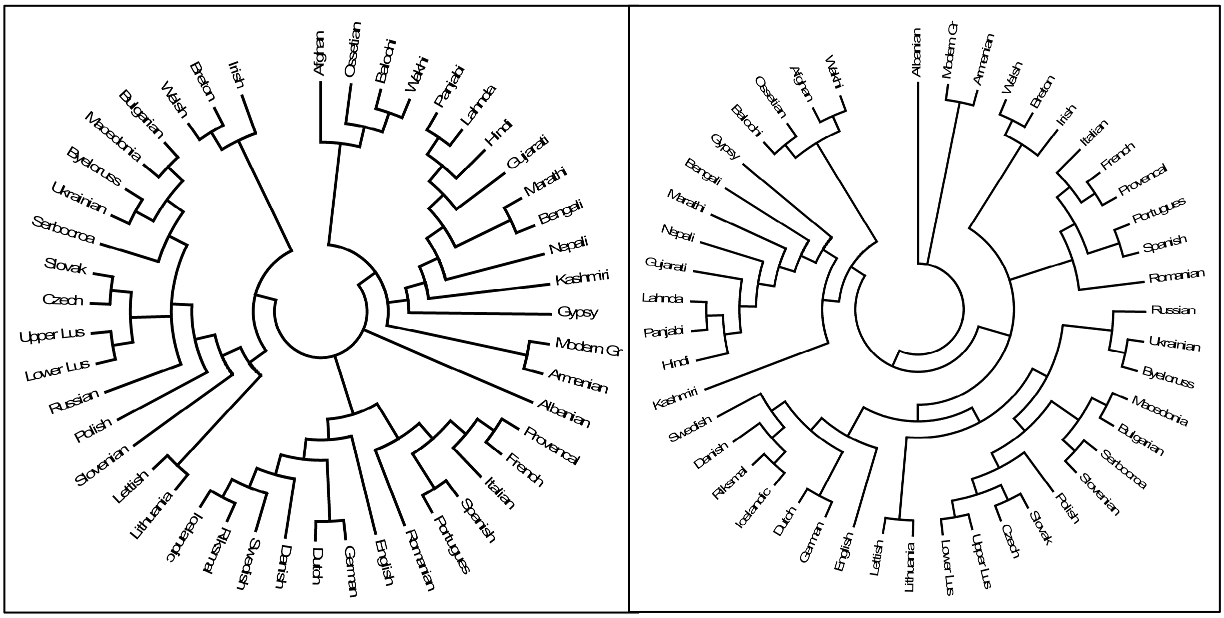
3.4. Comparison of "BIRD" in Tower of Babel (no date) and Dyen et al. (1997):

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

3.5. Undetected Borrowings in the Romance Partition in ToB and Dyen

	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			

3.6. The Tree Topologies of the Bayesian Analyses[4]



Dyen **ToB**

4. Back to the Roots

Rea (1973) on the Validity of Lexicostatistics:

- "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 indeterminencies, inconsistencies and errors, what then remains?" (Rea 1973: 361)

Root-Based Analyses which have been Proposed so far:

Holm 2000, 2005, 2008	Separation Base Method: Estimating genetic distances between languages using a hypergeometrical estimation of the root-size of ancestor languages based on etymological dictionaries
Starostin 2000	Etymostatistics: Estimating the genetic distance of languages by comparing the roots found in various texts of a certain language with the number of roots reflected in other genetically related languages
Ellegård 1959	Method similar to that proposed by Holm (2001, 2005, 2008), but with a different formula for data normalization

Plans for the Future within our Research Project

- Testing root-based approaches
- Biology and linguistics: Investigation of transferability of methods and theories
- Making the methods more scientific: Increasing transparency and the quality of the data

[4] Analysis was made using MrBayes (cf. Ronquist et al. 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),

Literaturverzeichnis

- Atkinson, Q. D.& Gray, R. D. (2006): How Old is the Indo-European Language Family? Illumination or More Moths to the Flame? In: Forster, P.&Renfrew, C. (Hgg.): Phylogenetic methods and the prehistory of languages. Cambridge UK , Oxford UK , Oakville CT USA ,. 91–109.
- Atkinson, Q. D.; Meade, A.; Venditti, Chris& Pagel, M. D. (2008): Languages evolve in punctuational bursts. *Science.* 319. 5863. 588.
- Atkinson, Q. D.; Nicholls, G. K.; Welch, D.& Gray, R. D. (2005): From words to dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society.* 103. 2. 193–219.
- Bergsland, K.& Vogt, H. (1862): On the Validity of Glottochronology. *Current Anthropology.* 3. 115-153.
- Blust, R. (2000): Why lexicostatistics doesn’t work. The 'universal constant' hypothesis and the Austronesian languages. In: Renfrew, C.; McMahon, A.&Trask, L. (Hgg.): Time depth in historical linguistics. Cambridge. 311–331.
- Burlak, S. A.& Starostin, S. A. (2005): Sravnitel'no-istoričeskoe jazykoznanie. Moskva. Ucebnoe izd.
- Dyen, I.; Kruskal, J. B.& Black, P. (1992): An Indoeuropean classification : a lexicostatistical experiment. Philadelphia.
- Dyen, I.; et al.: Comparative Indoeuropean database collected by Isidore Dyen. *Letzte Aktualisierung:* Mittwoch, 5. Februar 1997. *Zuletzt geprüft am:* Donnerstag, 26. Februar 2009. http://www.wordgumbo.com/ie/cmp/iedata.txt.
- Ellegård, A. (1959): Statistical Measurement of Linguistic Relationship. *Language.* 35. 2(Part 1). 131–156.
- Gray, R. D.& Atkinson, Q. D. (2003): Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature.* 426. 6965. 435–439.
- Greenhill, S. J.; Blust, R.& Gray, R. D. (2008): The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics.* 4. 271–283.
- Hoenigswald, H. M.&Langacre, R. H.(Hgg.)(1973): Diachronic, areal and typological linguistics. The Hague; Paris.
- Hoijer, H. (1956): Lexicostatistics. A Critique. *Language.* 32. 1. 49–60.
- Holm, H. J. (2007): The Distribution of Data in Word Lists and its Impact on the Subgrouping of Languages. :. In: Preisach, C.; et al. (Hgg.): Data Analysis, Machine Learning, and Applications. Proc. of the 31th Annual Conference of the German Classification Society (GfKl). Heidelberg-Berlin. 629–636.
- [Der Titel kann nicht dargestellt werden – Die Vorlage "Literaturverzeichnis - - (Standardvorlage)" enthält keine Informationen.]
- Lees, R. B. (1953): The Basis of Glottochronology. *Language.* 29. 2. 113–127.
- Matisoff, J. A. (1978): Variational semantics in Tibeto-Burman. The 'organic' approach to linguistic comparison. Philadelphia.
- Matisoff, J. A. (2000): On 'Sino-Bodic' and Other Symptoms of Neosubgroupitis. *Bulletin of the School of Oriental and African Studies.* 63. 3. 356–369.
- McMahon, A.& McMahon, R. (2005): Language classification by numbers. Oxford.
- McMahon, A.& McMahon, R. (2006): Why Lingustics Don't Do Dates: Evidence from Indo-European and Australian Languages. In: Forster, P.&Renfrew, C. (Hgg.): Phylogenetic methods and the prehistory of languages. Cambridge UK , Oxford UK , Oakville CT USA ,. 153--160.
- Meiser, G. (1999): Historische Laut- und Formenlehre der lateinischen Sprache.
- Milke, W. (1962): "Comment" on Bergsland & Vogt (1962).
- Pagel, M. D.; Atkinson, Q. D.& Meade, A. (2007): Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature.* 449, 7163. 717–721.
- Rea, J. A. (1973): The Romance data of pilot studies for glottochronology. In: Hoenigswald, H. M.&Langacre, R. H. (Hgg.): Diachronic, areal and typological linguistics. 11. The Hague; Paris. 355–367.
- Rexová, K.; Bastin, Y.& Frynta, D. (2006): Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften.* 93. 4. 189–194.
- Ronquist, F.& Huelsenbeck, J. P. (2003): MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.* 19. 12. 1572–1574.
- Sagart, L.& Lee, Y.-J. (2008): No limits to borrowing. The case of Bai and Chinese. *Diachronica.* 25. 3. 357–385.
- Searls, D. B. (2003): Linguistics: Trees of life and of language. *Nature.* 426. 6965. 391–392.
- Starostin, S. A. (1993): Rabočaja sreda dlja lingvista (Working environment for a linguist): Bazy dannyh po istorii Evrazii v srednie veka. 7–23.
- Starostin, S. A. (2000): Comparative-historical linguistics and lexicostatistics. In: Renfrew, C.; McMahon, A.&Trask, L. (Hgg.): Time depth in historical linguistics. Cambridge. 223–265.

- Swadesh, M. (1950): Salish Internal Relationships. *International Journal of American Linguistics.* 16. 4. 157–167.
- Swadesh, M. (1952): Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society.* 96. 4. 452–463.
- Swadesh, M. (1955): Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics.* 21. 2. 121–137.
- Swadesh, M. (1962): "Comment" in Bergsland & Vogt (1962).
- The Tower of Babel. An Etymological Database Project. http://starling.rinet.ru/main.html.
- Tischler, J.& Ganter, B. (1997): Review of I. Dyen, J. Kruskal & P. Black: An Indoeuropean Classification (1992). *Kratylos.* 42. 43–50.
- Wikipedia: Wiktionary. *Letzte Aktualisierung:* Sonntag, 20. September 2009. *Zuletzt geprüft am:* Mittwoch, 23. September 2009. http://de.wikipedia.org/w/index.php?title=Wiktionary&oldid=64738396.

Appendix: Making the Data Comparable

- in order to make the datasets comparable, 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
- both loans and gaps are coded by assigning negative numbers to the words
- additionally, 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)
- cognate judgments were restricted 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 this way, we replaced all numbers which would show up in different rows of items by new numbers)