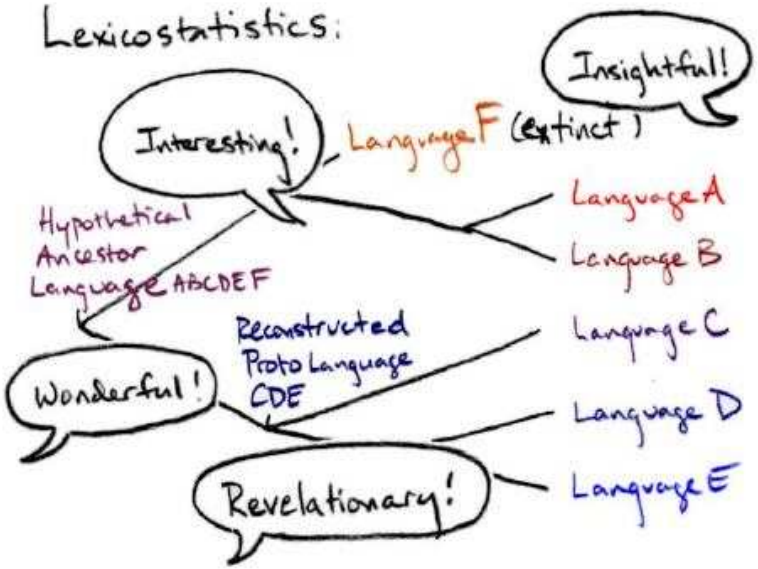
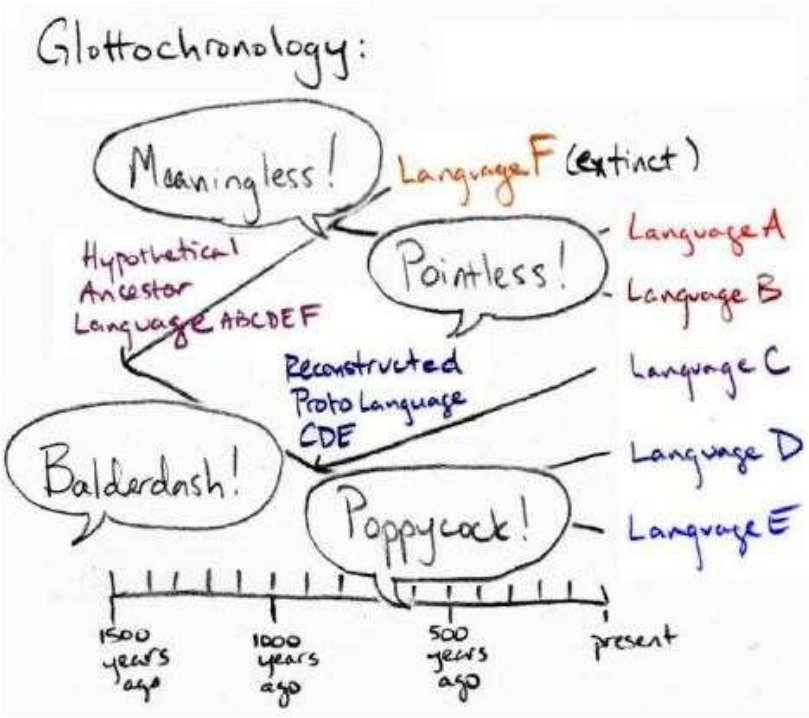
Schöne Bäume auf schwankendem Grund

Zum Datenproblem in der Lexikostatistik

Hans Geisler & Johann-Mattis List





Einleitung

LEXIKOSTATISTIK

Lexikostatistik

Theoretische Grundannahmen:

- The lexicon of every human language contains words which express universal concepts, 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

Lexikostatistik

Vorgehensweise (theoretisch)

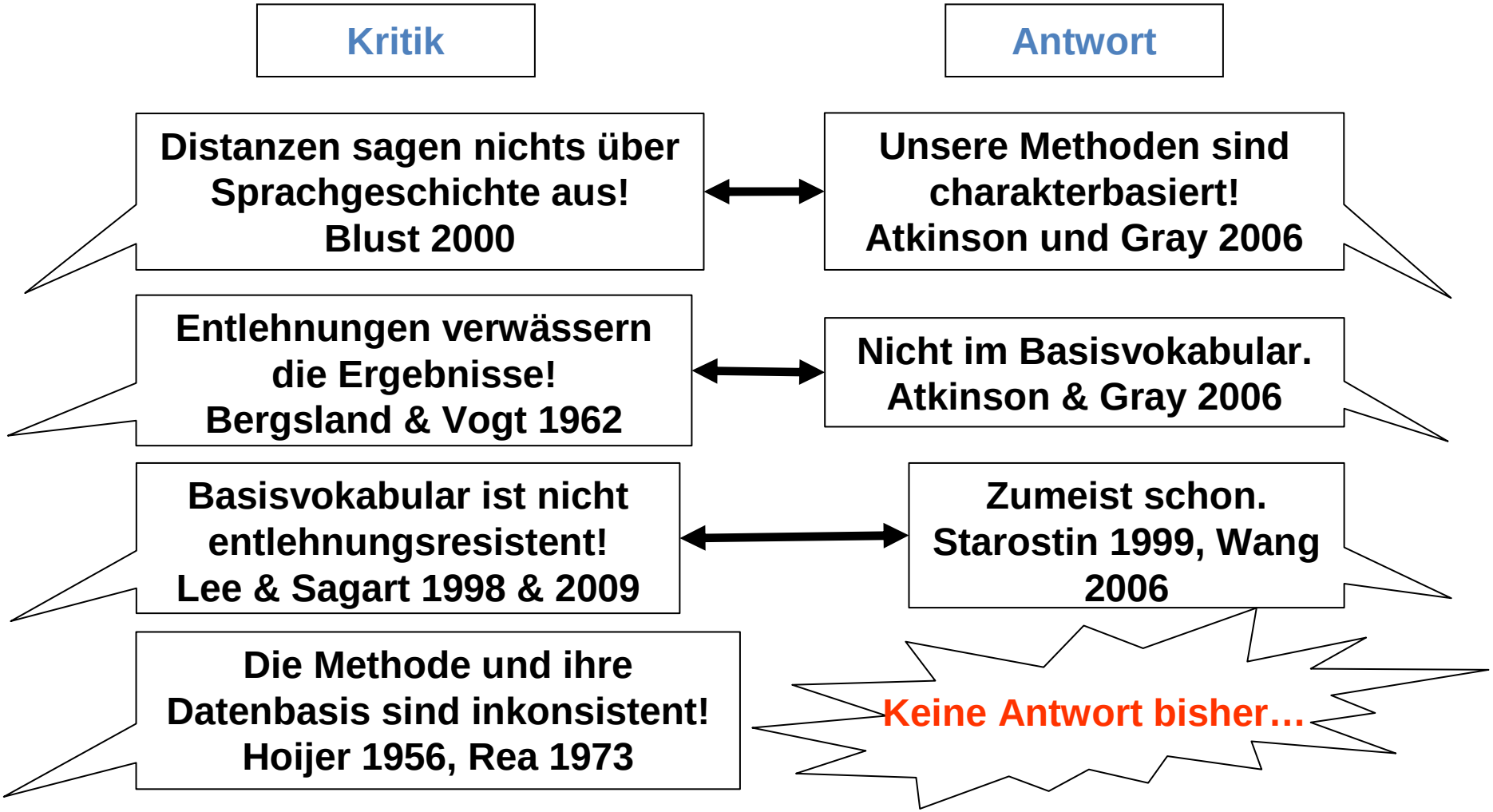
1	Basisvokabular wählen	Bedeutungsliste erstellen (oder eine von ca. 40 bisher postulierten auswählen)
2	Wortlistenerstellung	Wörter für die jeweiligen Bedeutungen in die Einzelsprachen, die untersucht werden sollen, übersetzen
3	Kognatenzuweisung	Etymologisch verwandte Wörter in den Einzelsprachen mit Hilfe der komparativen Methode bestimmen
4	Kodierung	Daten numerisch aufbereiten
5	Analyse	Numerisch aufbereitete Daten in ein graphisches Darstellungs-Format überführen (meistens Bäume)

Lexikostatistik

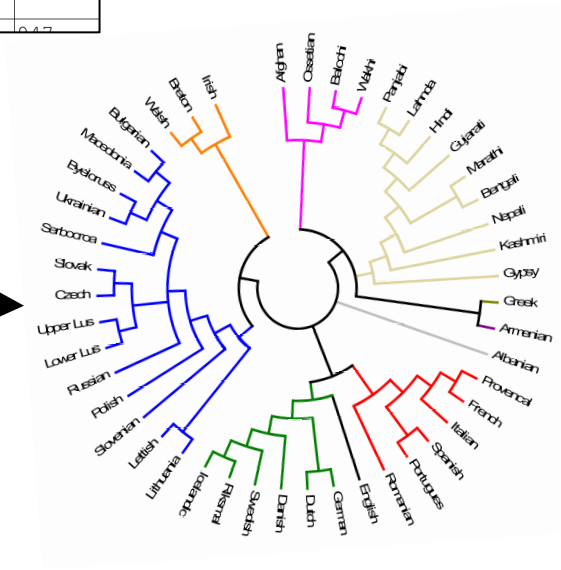
Vorgehensweise (praktisch)

1	Basisvokabular wählen	Nimm Swadesh-100, Starostin-110, Wiktionary-207 oder erstelle eine eigene, intuitiv plausible Bedeutungsliste.
2	Wortlistenerstellung	Schau nach, ob die Daten im Internet zu finden sind, ansonsten nimm ein zweisprachiges Taschen-wörterbuch und übersetze die Bedeutungen in die Zielsprache.
3	Kognatenzuweisung	Einfach auf die Intuition verlassen und (ab und zu mal im Pokorny nachschauen).
4	Kodierung	Überführe die Kognaten in ein numerisches System, oder frage einen Biologen oder Mathematiker.
5	Analyse	Erstelle eine Klassifikation mit phylogenetischer Software oder frage einen Biologen oder Statistiker.

Hauptkritikpunkte



12	burn	-	-1	-	-1
15	cold	friguros	1102	freddo	1102
15	cold	rece	1520	-	0
20	dry	sec	2390	secco	2390
20	dry	uscat	3207	asciutto	3207
22	earth	-	0	terra	1831
22	earth	pământ	1977	-	0
22	earth (soil)	sol	9003	suolo	9003
26	fat(n)	grăsime	1539	grasso	1539
26	fat(n)	-	0	-	0



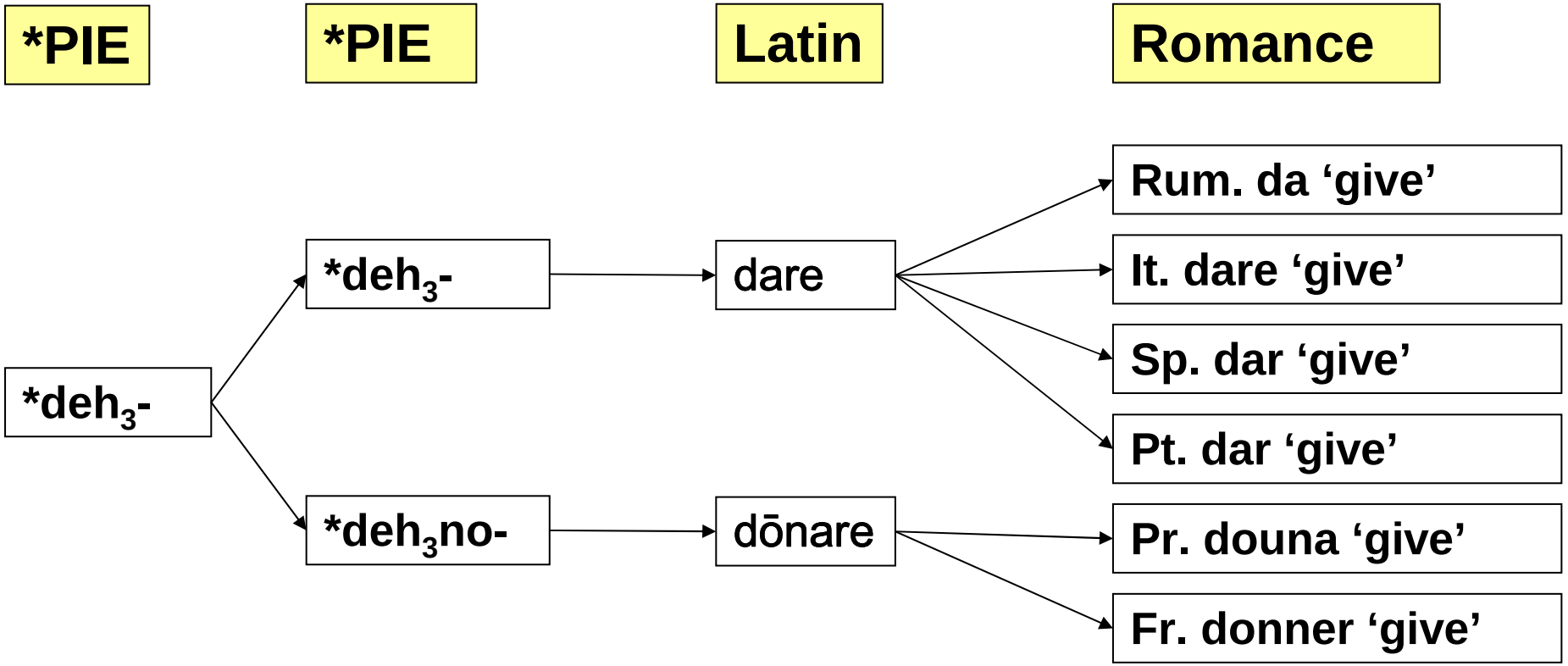
Teil I

Datenprobleme

Wortlistenerstellung (Schritt 2)

- methodenbedingte Fehler
  - Konzeptunschärfen
  - Synonymendifferenzierung
  - Varianz (diastratisch, diatopisch, etc.)
- bearbeiterbedingte Fehler
  - Mangelnde Kompetenz in der Einzelsprache
  - Verwendung minderwertiger Quellen

Kognatenzuweisung (Schritt 3)



hand	Hand
head	Kopf
kill	killen
sun	Sonne
fat	Fett
short	kurz

Denksportaufgabe für angehende Lexikostatistiker: Finden Sie die Kognaten und die Lehnwörter!

Teil II

Listenvergleich

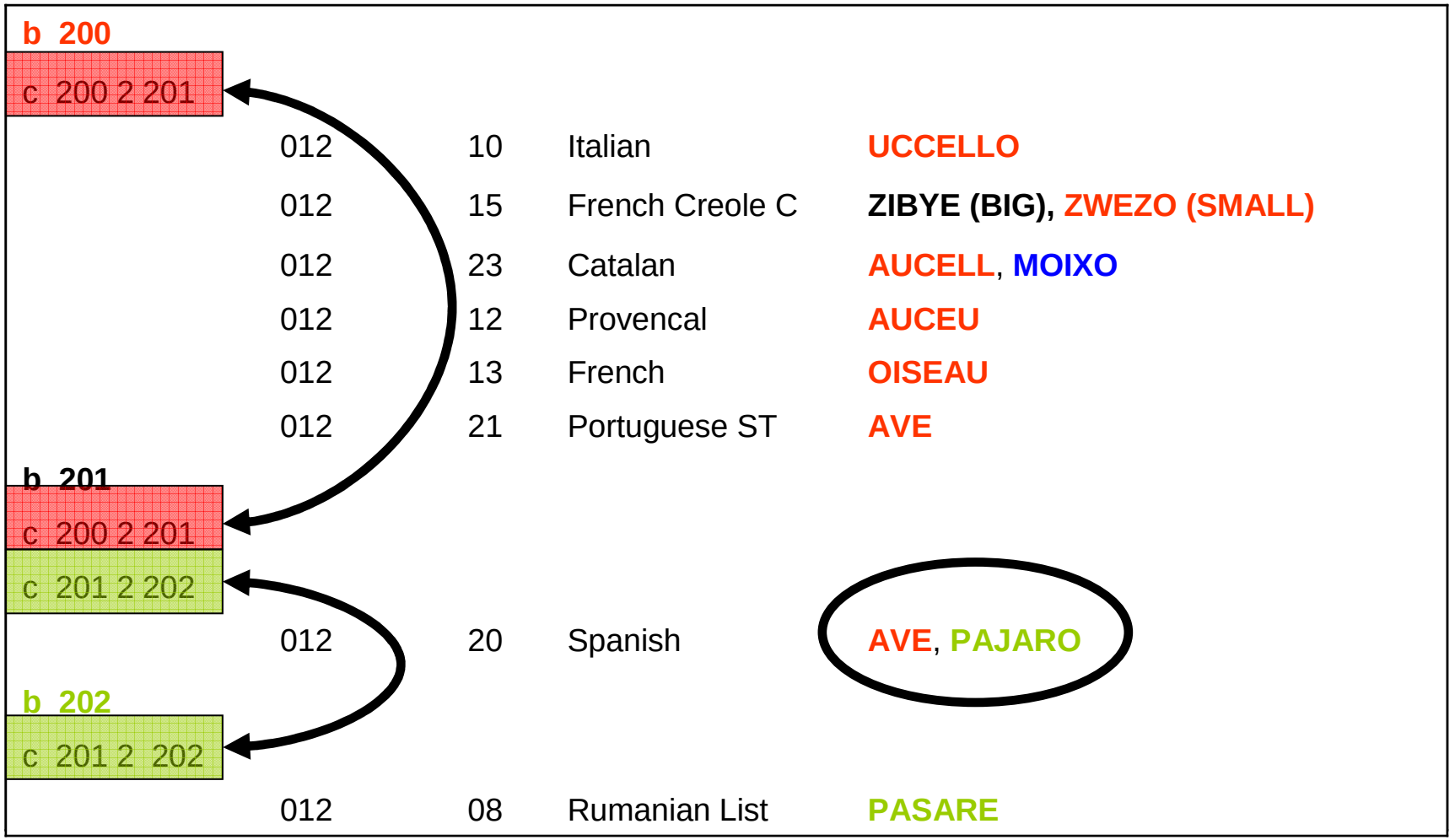
Listenvergleich

Vergleich der Einträge in zwei unabhängig voneinander erstellten Swadesh-Listen

Autor	Dyen , Kruskal & Black (1997)	Tower of Babel (o.J.)	Schnittmenge
Sprachfamilie	Indogermanisch	Indogermanisch	Indogermanisch
Anzahl von Spr.	95	98	46
Anzahl von Items	200	110	103

Listenvergleich

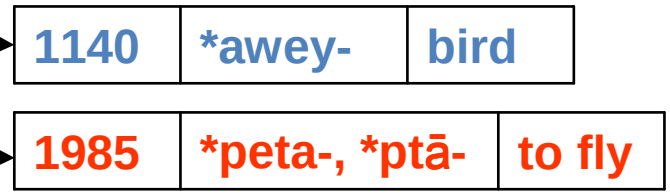
Dyen et al (1997): BIRD



Listenvergleich

Tower of Babel (o. J.): 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**		



Listenvergleich

Tower of Babel (o. J.) vs. Dyen et al. (1997):

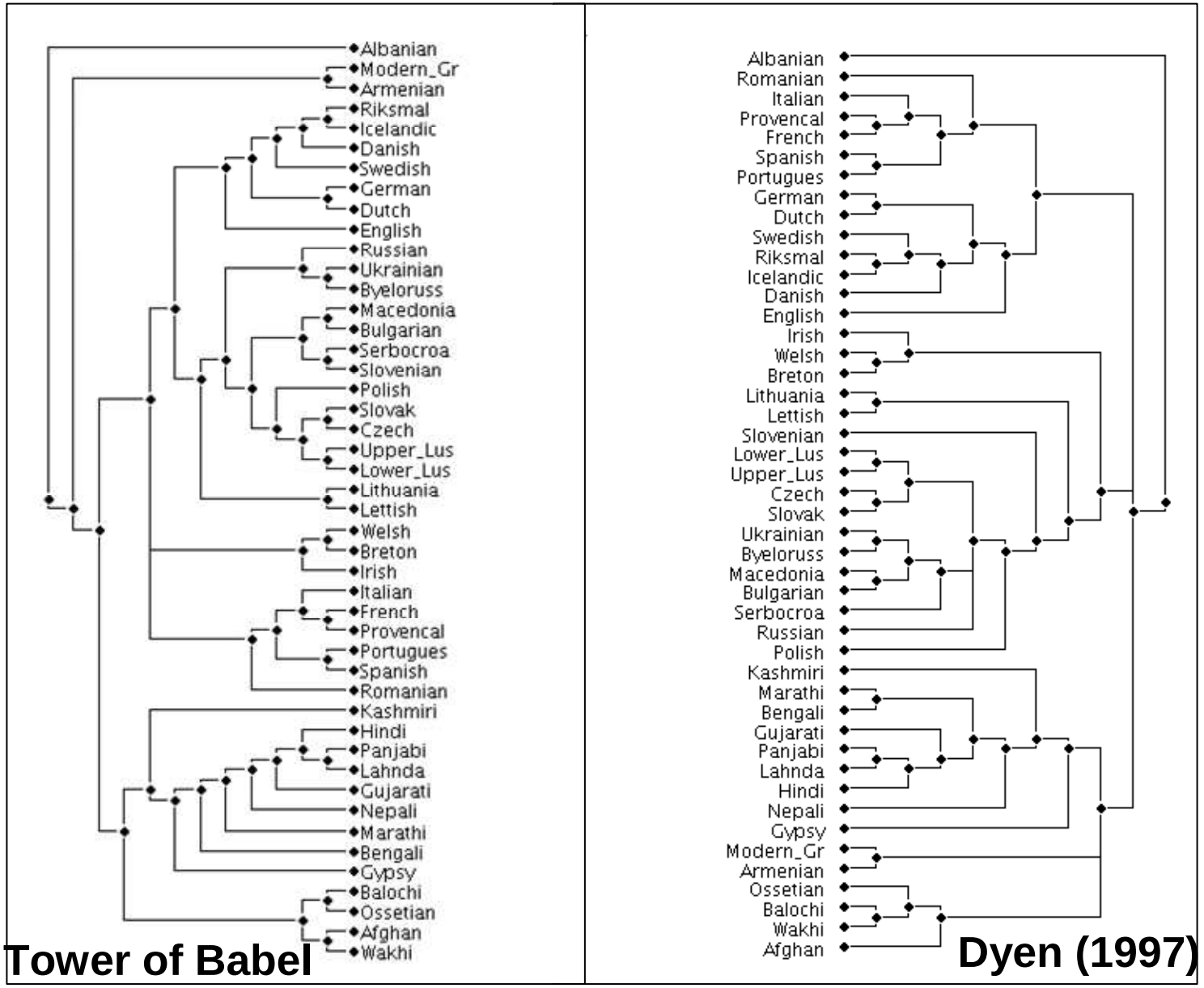
BIRD	*Dyen*	*ToB*		*G&L*	
it.	UCCELLO	uccello		uccello	passero
fr.	OISEAU	oiseau		oiseau	passereau
pt.	AVE	ave	passaro	ave	pássaro
sp.	AVE, PAJARO	ave	pajaro	ave	pájaro
pr.	AUCEU	aucel		aucel	paser
rum.	PASARE		pasăre		pasăre

Listenvergleich

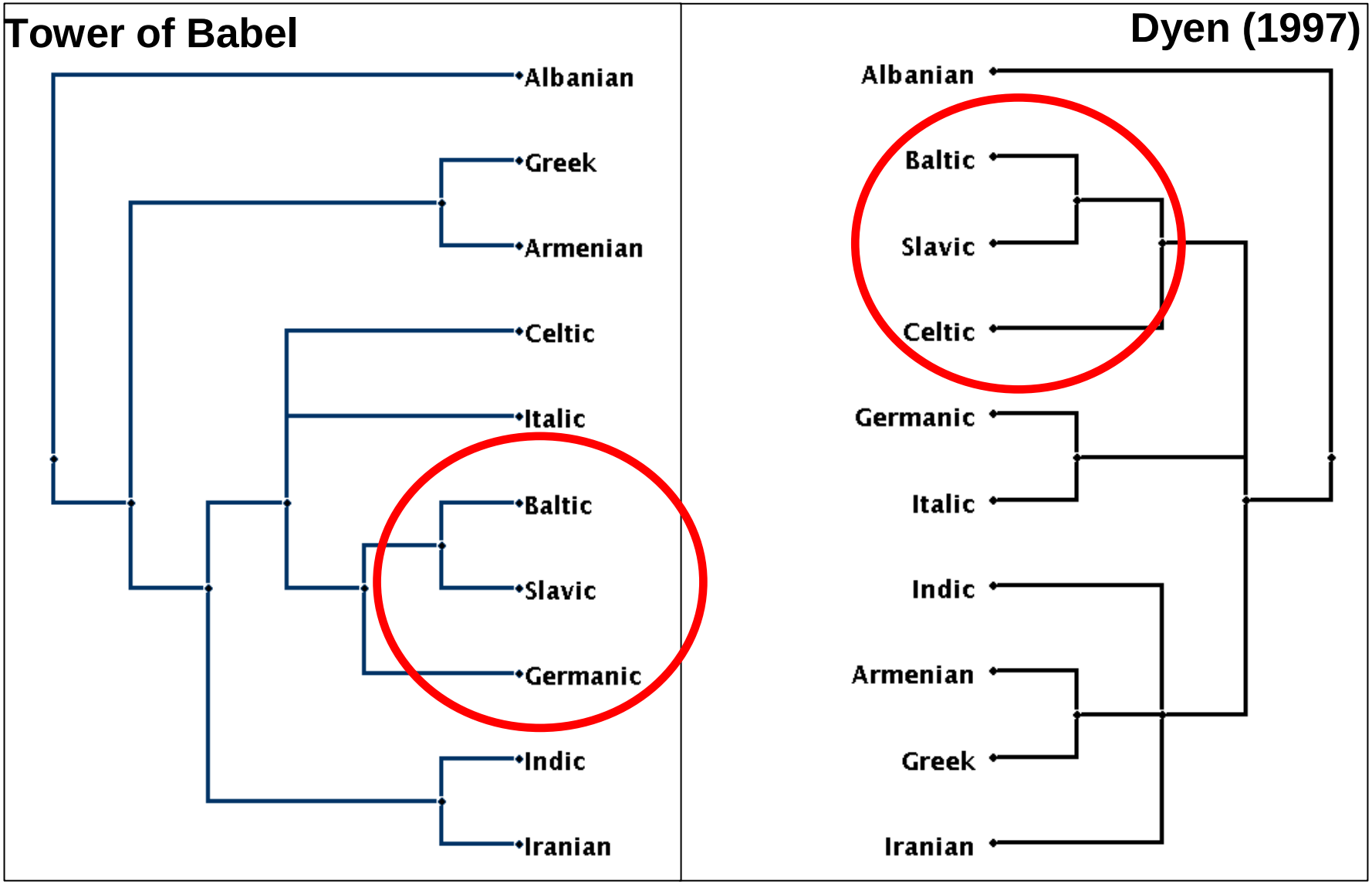
Nicht entdeckte Entlehnungen

	Item	Donor	Quelle	rum.	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**			

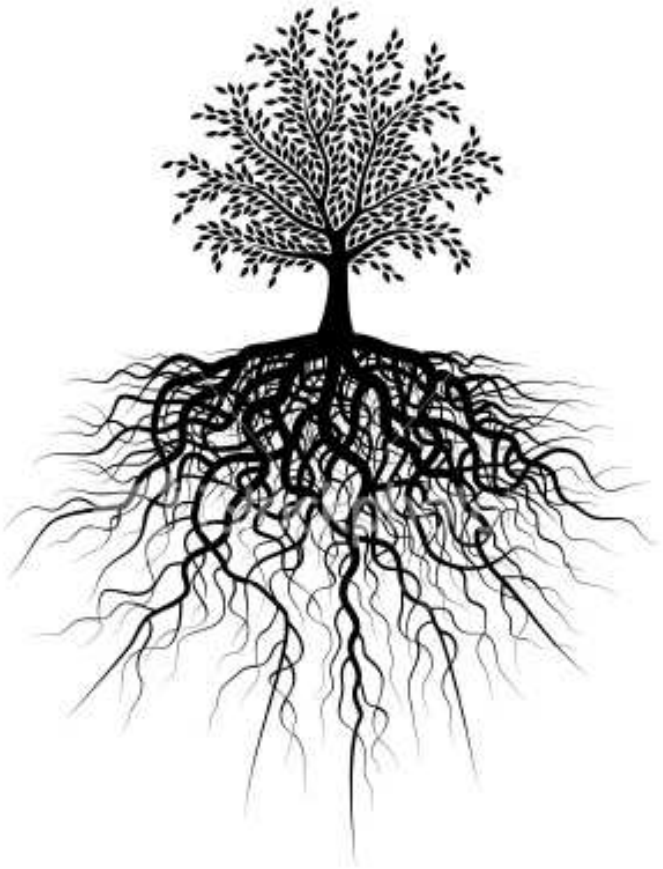
Die schönen Bäume für bayesianische Analysen



Vergleich der Supergruppen in den bayesianischen Analysen



> **“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)**



Schluss

Back to the roots…

Das Datenproblem

- Ein konsistentes Übersetzen der lexikostatistischen Bedeutungslisten in die Einzelsprachen ist aufgrund der semantischen Variation innerhalb dieser nicht möglich und führt stets zu einer subjektiven Vorauswahl der Daten
- Wenn die Bedeutungslisten an individuelle Sprachfamilien angepasst werden, um der semantischen Variation gerecht zu werden, verlieren die lexikostatistischen Grundannahmen bezüglich des Basisvokabulars (Universalität, Stabilität, Resistenz) ihre Gültigkeit
- Wenn die lexikostatistischen Grundannahmen bezüglich des Basisvokabulars nicht mehr zutreffen, verliert auch das zweite Postulat der Lexikostatistik (dass geteilte Retentionen innerhalb des Basisvokabulars Aussagen über den Verwandtschaftsgrad von Sprachen zulassen) seine Gültigkeit
- Jede „lexikostatistische“ Klassifikation ist daher willkürlich und subjektiv und folglich auch nicht valide

Wurzelbasierte Ansätze

- Während lexikostatistische Ansätze Kognazitätsurteile von „Bedeutungsgleichheit“ abhängig machen, wird Kognazität in wurzelbasierten Ansätzen im Rahmen der komparativen Methode definiert (semantische Identität ist kein Kriterium für das Postulieren von Kognaten)
- Die Abkehr von den semantischen Restriktionen macht es möglich, größere Datensätze für die Analyse zu verwenden
- Die Hinwendung zur komparativen Methode für die Erstellung von quantitativen Datensätzen macht diese Ansätze (hoffentlich) wissenschaftlicher und objektiver

Zurück zu den Wurzeln…

Einige Punkte, denen wir im Rahmen unseres Projektes nachgehen wollen

- Testen wurzelbasierter Ansätze (Starostin 1989/2000, Holm 2001 & 2007, Ellegård 1959, Herdan 1966)
- Verwissenschaftlichung der Methoden: Steigerung der Transparenz und der Qualität der Datenbasis, Formalisierung der Arbeitsprozesse
- Evolutionsbiologie und Linguistik: Untersuchung der Übertragbarkeit von Konzepten und Methoden zwischen den Disziplinen

BMBF-Förderung

- Förderschwerpunkt
  - *Wechselwirkungen zwischen Natur- und Geisteswissenschaften*
- Thema
  - *Klassifikation und Evolution in Biologie, Linguistik und Wissenschaftsgeschichte*
- Interdisziplinäre Forschergruppe
  - *Heiner Fangerau (Wissenschaftsgeschichte, Univ. Ulm)*
  - *William Martin (Genetik, HHU Düsseldorf)*
  - *Hans Geisler (Linguistik, HHU Düsseldorf)*
- Laufzeit
  - *2009-2011*

Evolution of Language

- "There is perhaps no field of scientific study in which more progress has been made—in spite of a complete lack of any clear information on which to base either theories or conclusions—than in the study of the evolution of human language. The pioneers in this arduous endeavor are to be highly commended for their intrepid tackling of a task of unparalleled difficulty, and for the amazing progress they have made, in spite of having no shoulders (or linguistic data) on which to stand." (Merritt Greenberg & Joseph Ruhlen, n.d.)