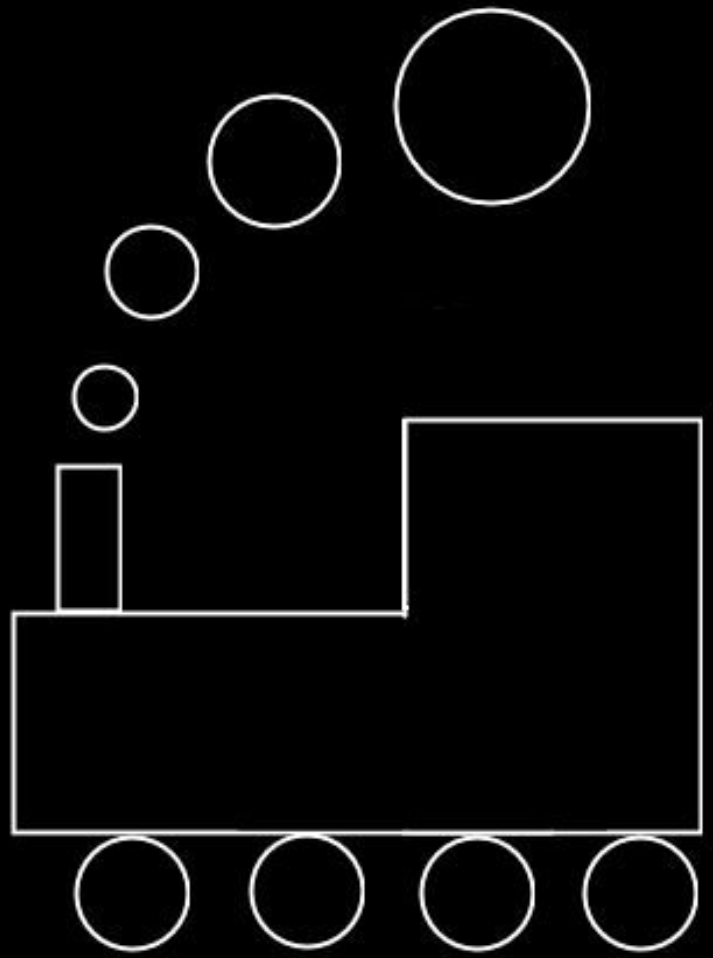
Die quantitative Wende in der historischen Linguistik: Chancen und Herausforderungen

Johann-Mattis List*

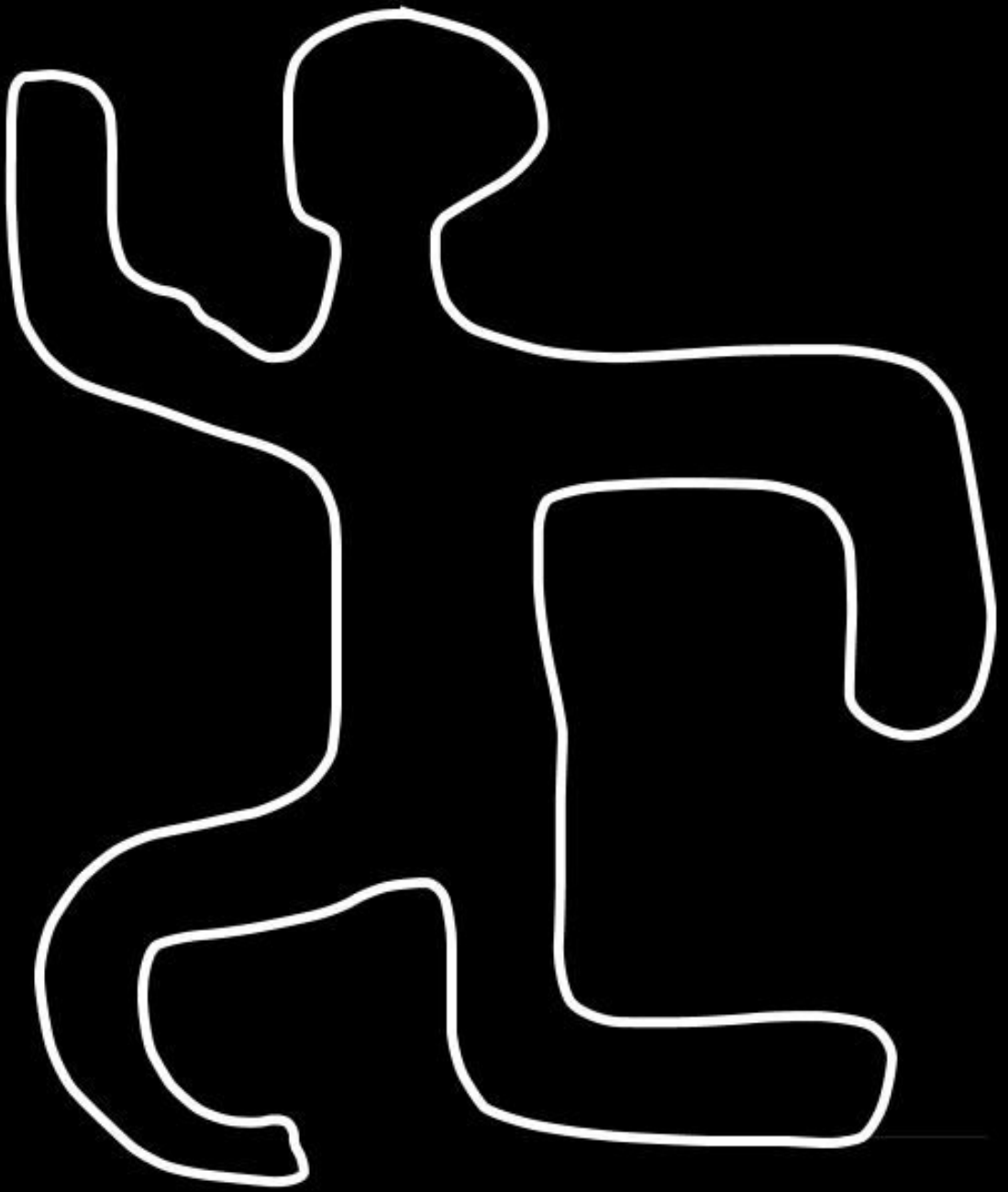
*Institut für Romanistik II
Heinrich Heine Universität Düsseldorf

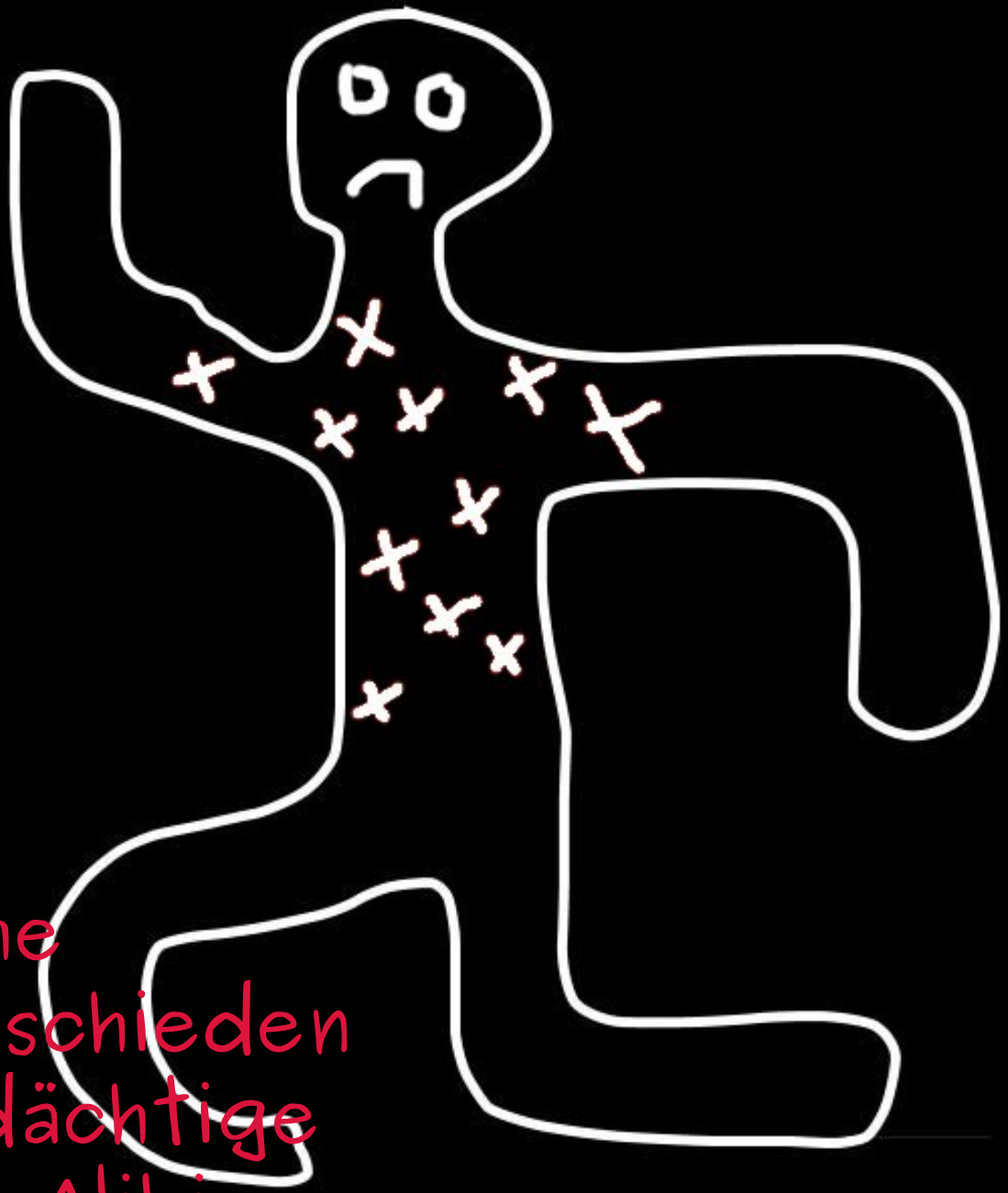
13. Mai 2012



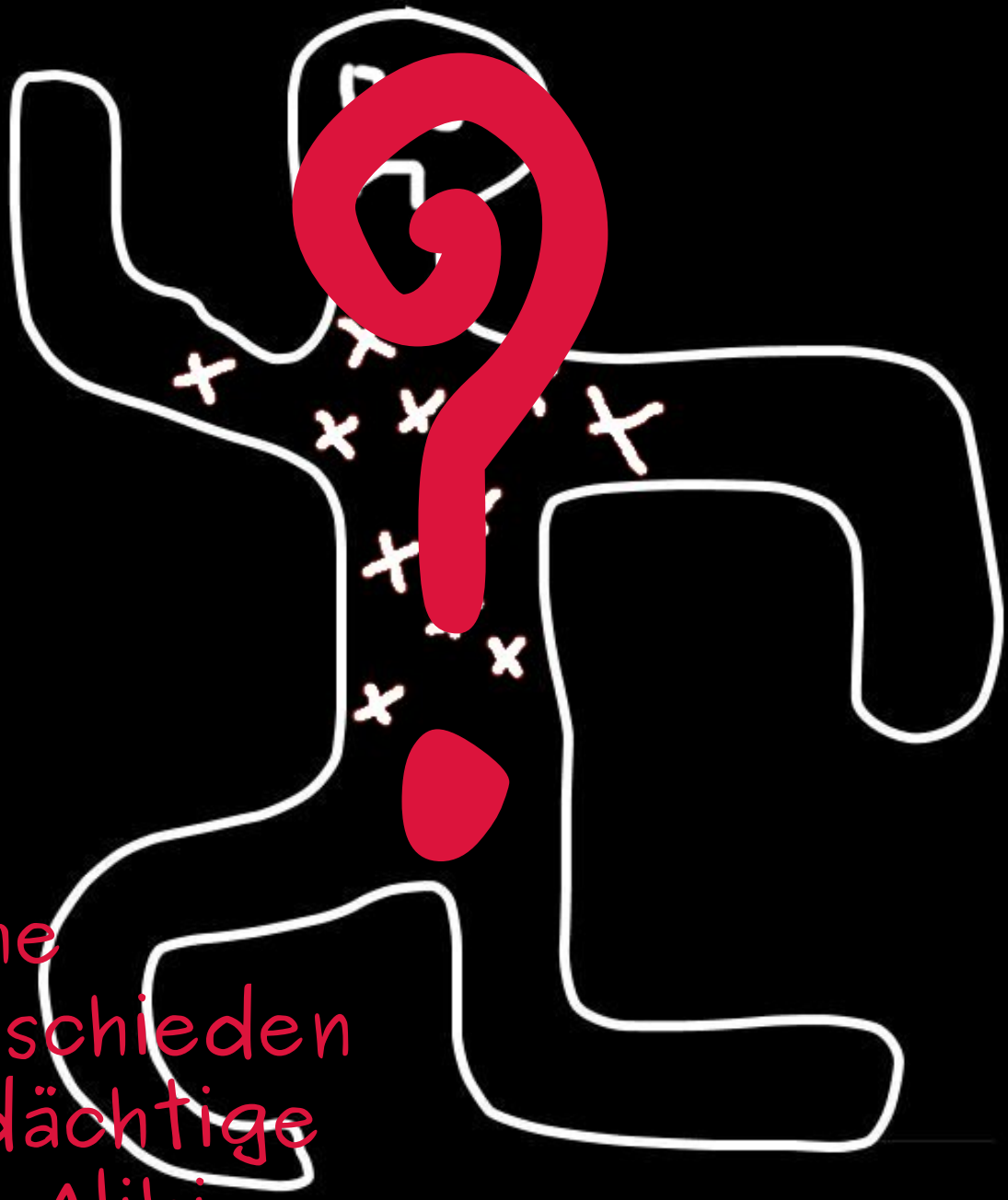
Mord im Orientexpress…

Mord im Orientexpress…





- 1 Toter
- 12 Stiche
- alle verschieden
- 12 Verdächtige
- alle ein Alibi



- 1 Toter
- 12 Stiche
- alle verschieden
- 12 Verdächtige
- alle ein Alibi

Was tun, Mr. Poirot?

Eh oui, Hastings, mir scheint, es gibt nur eine einzige Lösung, so abwegig sie auch scheinen mag: Es gab nicht einen einzigen Mörder, sondern gleich 12 davon…

Stimmt das, Mr. Holmes?

Well, why not? Meine Denkprozesse beruhen auf der Annahme, dass, wenn man alles ausgeschlossen hat, was unmöglich ist, das, was übrigbleibt, egal wie unwahrscheinlich es sein mag, die Wahrheit sein muss.

Alles schön und gut,
aber was soll das
jetzt eigentlich
mit historischer Linguistik
zu tun haben?

	Kriminalistik	Historische Linguistik
Ziel	den Mörder finden	die Ursprache finden
Vorgehen	Rekonstruktion des Tathergangs	Rekonstruktion der Sprachgeschichte
Methode	Indiziengestützte Beweisführung	Indiziengestützte Beweisführung

Kriminalistik

Historische Linguistik

*dent-

Agenda 2012

1. Traditionelle Historische Linguistik
   - Charakteristik
   - Errungenschaften
   - Probleme
2. Die quantitative Wende
   - Charakteristik
   - Errungenschaften
   - Probleme
3. Auf dem Weg zu einer qualitativen Wende?
   - Paradigmenwechsel
   - Beispiele
   - Ausblick

Traditionelle historische Linguistik

COMPENDIUM

DER

VERGLEICHENDEN GRAMMATIK

DER

INDOGERMANISCHEN SPRACHEN.

VON

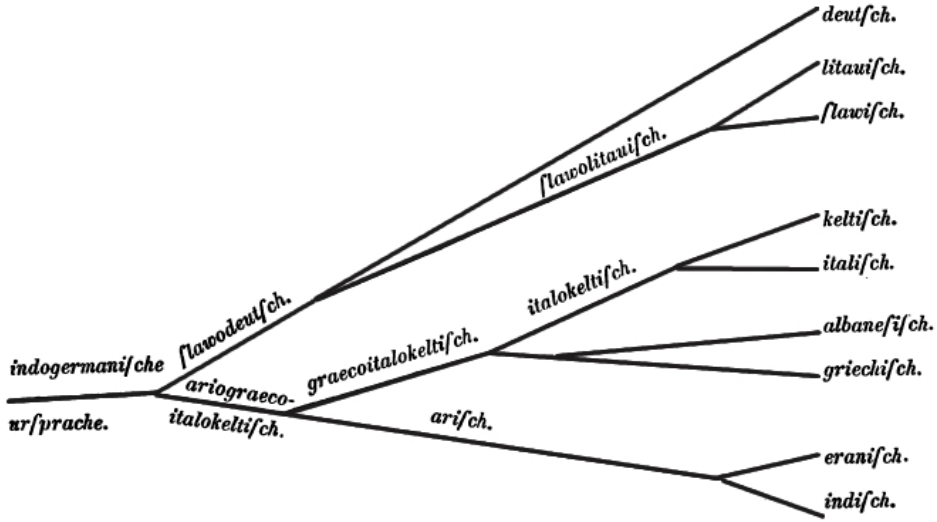
AUGUST SCHLEICHER.

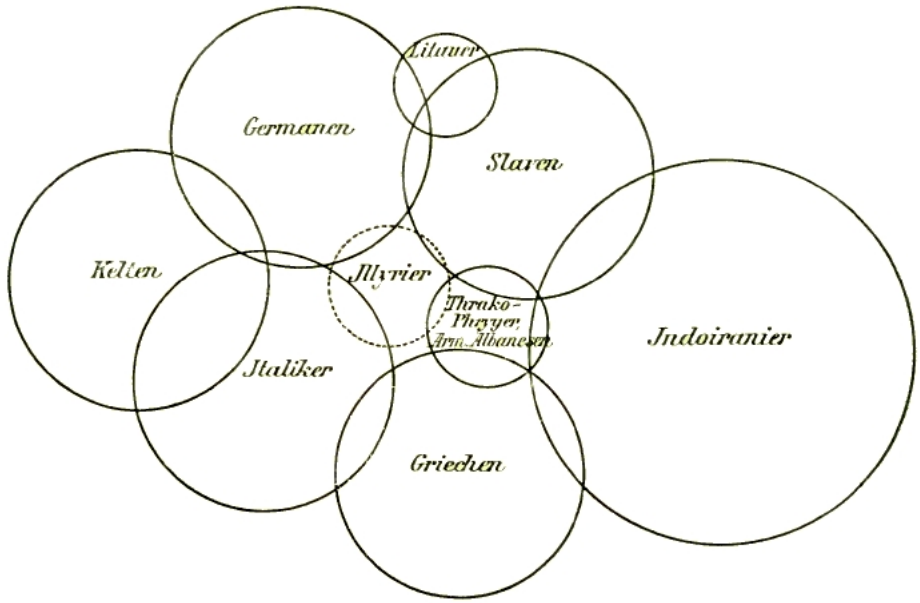
I.

WEIMAR

HERMANN BÖHLAU

1861.



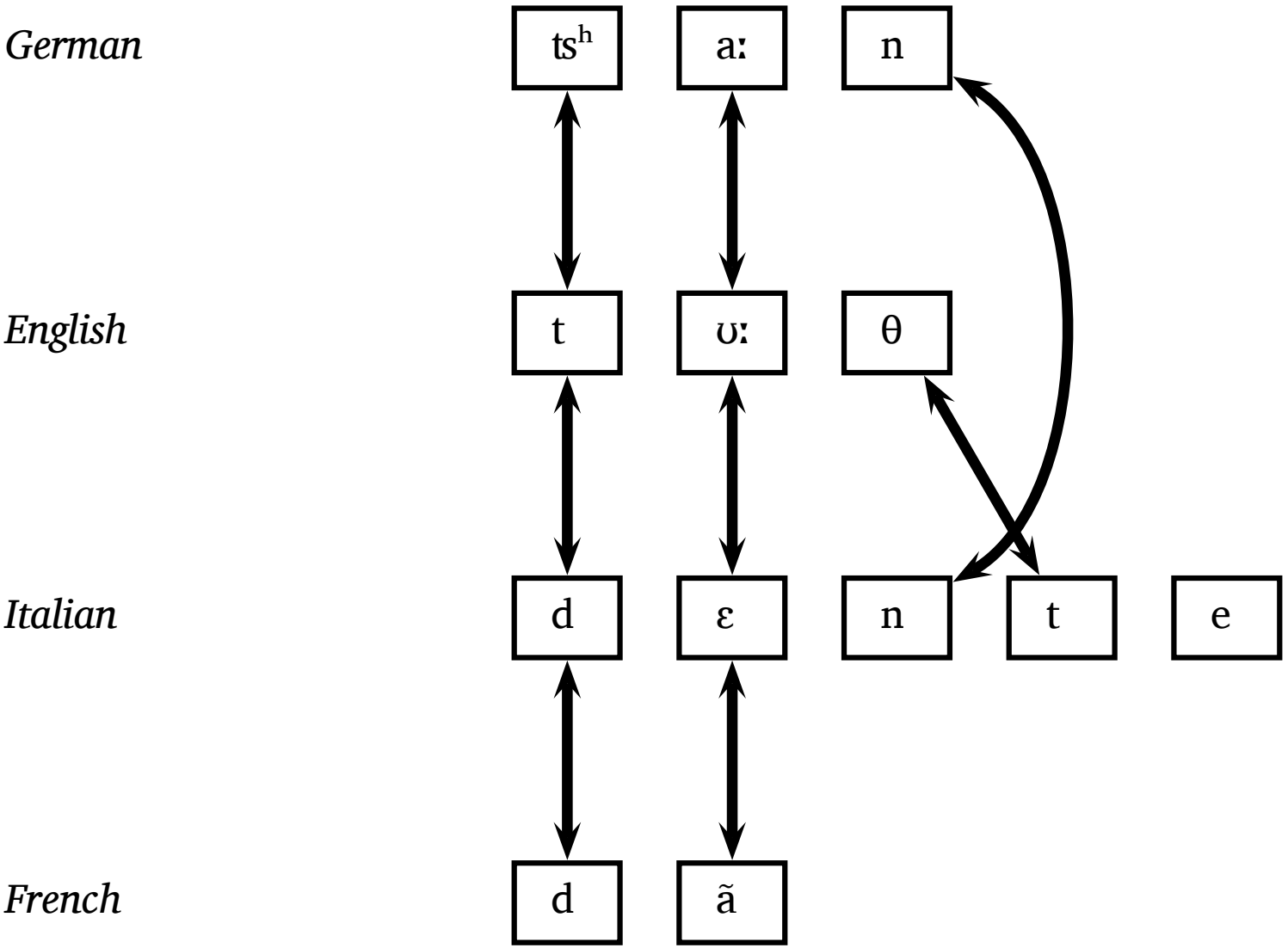


Charakteristik

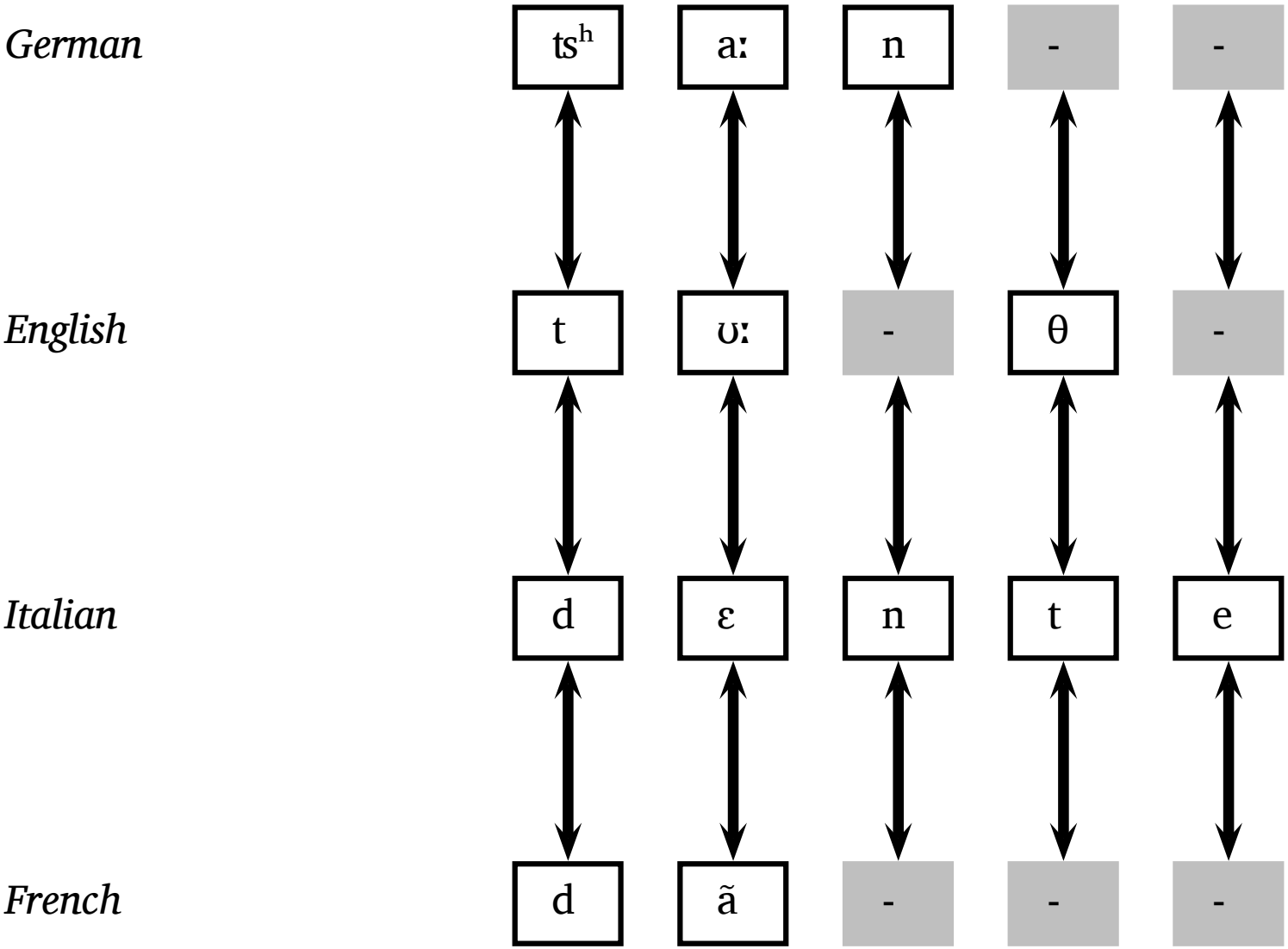
Forschungsgegenstand

German	tsʰ	aː	n		
English	t	ʊː	θ		
Italian	d	ɛ	n	t	e
French	d	ã			

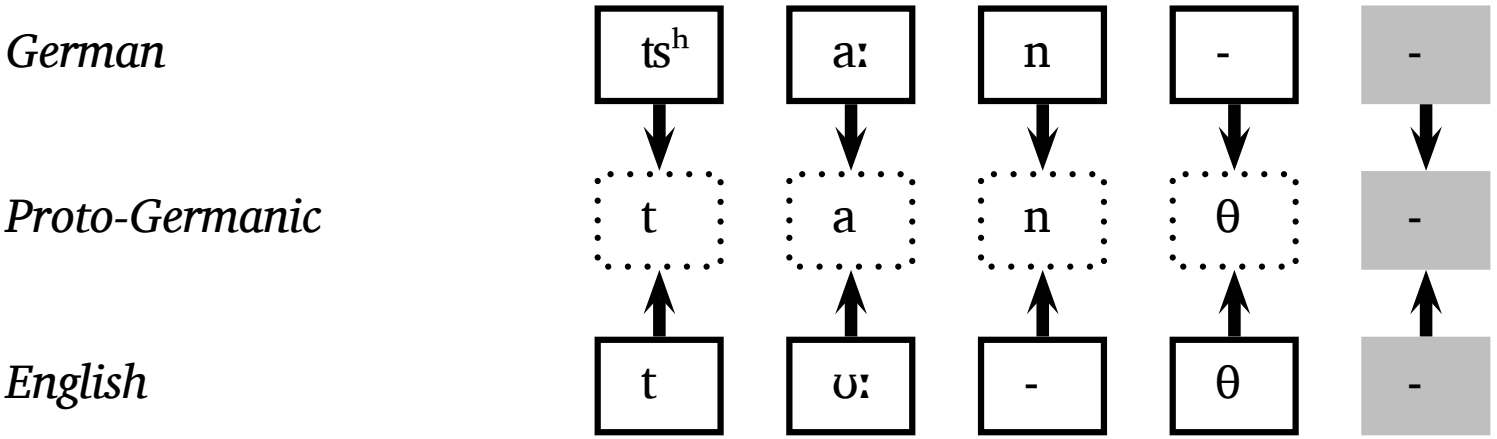
Forschungsgegenstand

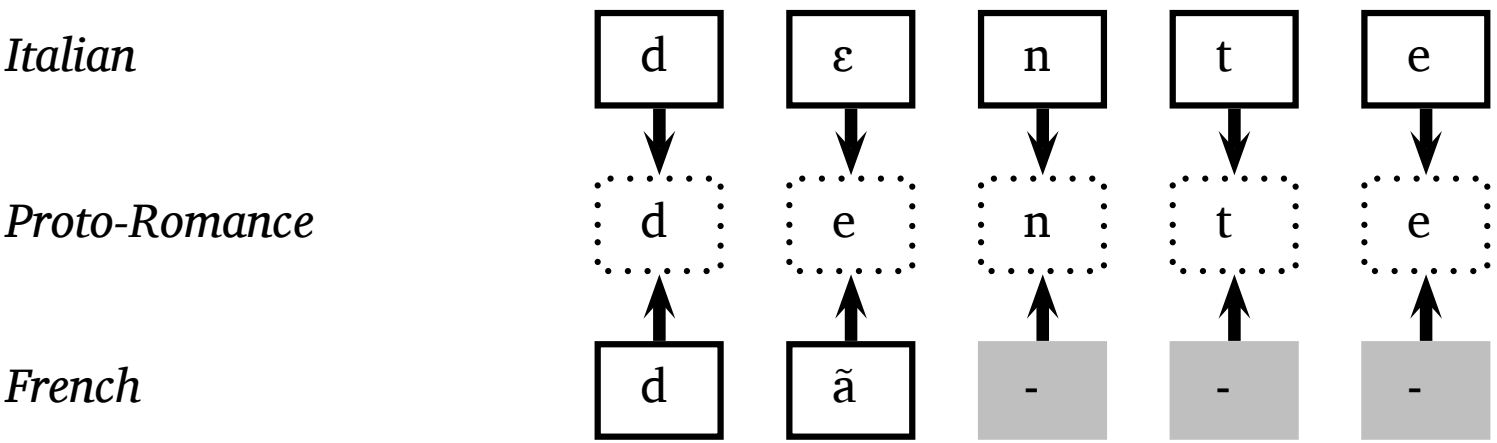


Forschungsgegenstand



Forschungsgegenstand

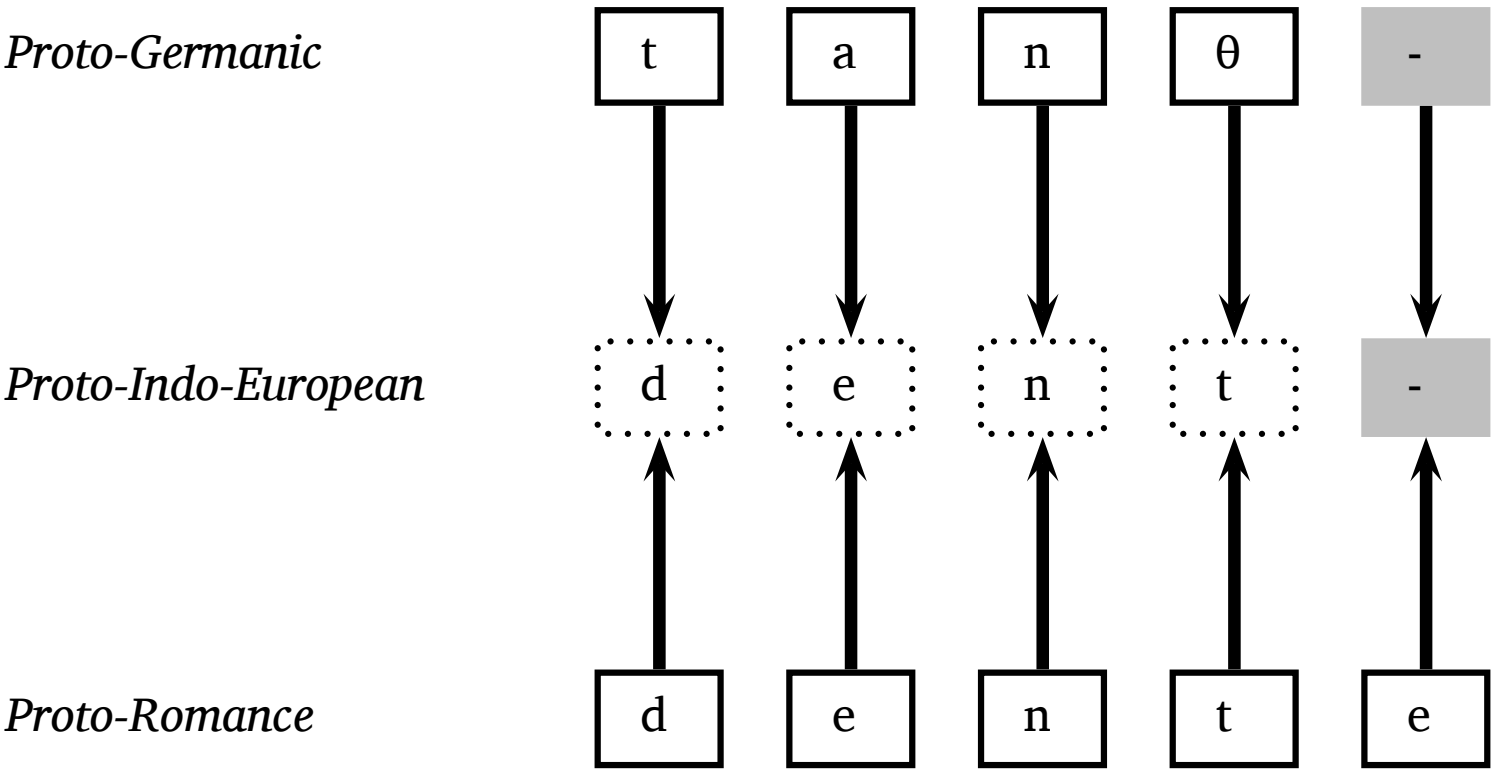




Forschungsgegenstand

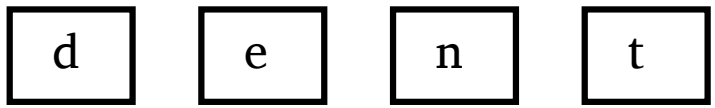
Proto-Germanic	t	a	n	θ	-
Proto-Romance	d	e	n	t	e

Forschungsgegenstand

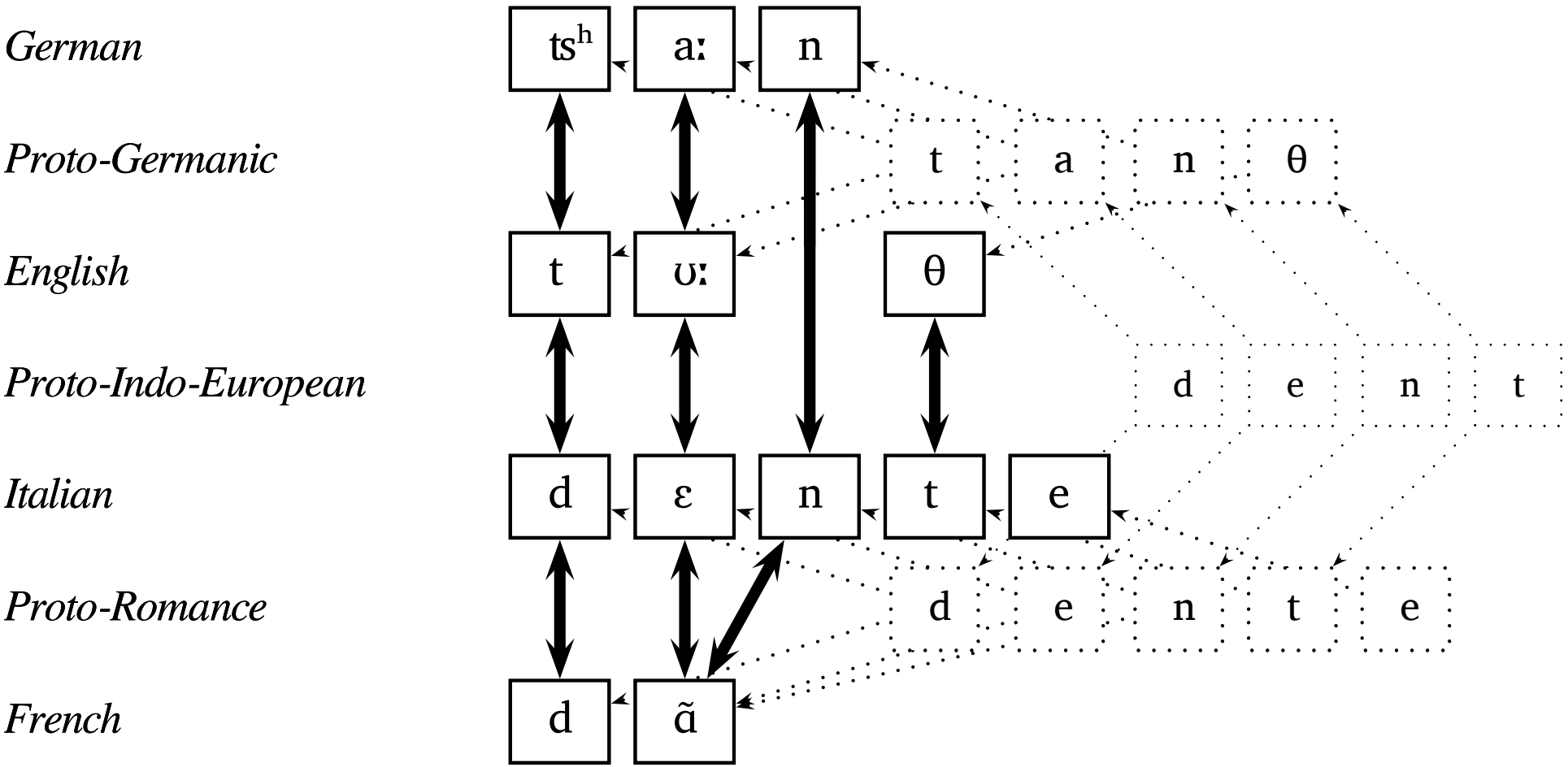


Forschungsgegenstand

Proto-Indo-European



Forschungsgegenstand



Forschungsgegenstand

Geschichte

- Individuelle Ereignisse (Beschreibung)
- Individuelle Prozesse (Beschreibung)
- Allgemeine Prozesse (Analyse)

Sprachgeschichte

- Individuelle Sprachzustände (Beschreibung von Lautsystem, Grammatik, Lexikon)
- Individuelle Sprachentwicklung (Beschreibung von Lautwandel, Grammatikalisierung, lexikalischem Wandel)
- Allgemeine Sprachentwicklung (Analyse von Prozessen des Lautwandels, der Grammatikalisierung, des lexikalischen Wandels)

Forschungsgegenstand

Innere Sprachgeschichte (Ontogenese)

- Etymologie
- historische Grammatik
- historische Phonologie

Äußere Sprachgeschichte (Phylogenese)

- linguistische Rekonstruktion
- Nachweise von Sprachverwandtschaft
- genetische Sprachklassifikation (phylogenetische Rekonstruktion)

Fragen der allgemeinen Sprachgeschichte

- Prozesse und Mechanismen des Lautwandels
- Grammatikalisierung
- lexikalischer Wandel

Ursprung

Uniformitarianismus

- “Universalität des Wandels” – Wandels verläuft unabhängig von Zeit und Raum
- “Gradualität des Wandels” – Wandel verläuft weder abrupt noch chaotisch
- “Uniformität des Wandels” – Wandel verläuft nicht heterogen, sondern einheitlich

Ursprung

Gründerväter

- Franz Bopp (1791–1867): Sprachvergleich (Bopp 1816)
- Rasmus Rask (1787-1832) und Jacob Grimm (1785-1863): Lautgesetz (Rask 1816, Grimm 1822)
- August Schleicher (1821–1868): Stammbaum und Rekonstruktion (Schleicher 1853 & 1861)

Errungenschaften

Methoden, Theorien und Modelle

Komparative Methode

Grundlegendes Verfahren zum Nachweis von Sprachverwandtschaft, zur linguistischen Rekonstruktion, zur Erstellung von Etymologien und zur genetischen Klassifikation

Stammbaummodell und Wellentheorie

Zwei (zum Teil widersprüchliche) Modelle zur Beschreibung von Verwandtschaftsbeziehungen zwischen Sprachen.

Regularitätshypothese

Bestimmte Lautwandelprozesse scheinen regelmäßig (universell, graduell und uniform) zu verlaufen.

Erkenntnisse

Innere Sprachgeschichte

Dank der historischen Linguistik ist eine beträchtliche (aber immer noch kleine) Anzahl von Sprachen hinsichtlich ihrer Entstehung sehr gut erforscht.

Äußere Sprachgeschichte

Dank der historischen Linguistik ist es gelungen, einen Großteil der Sprachen der Welt genetisch zu klassifizieren, wenn auch viele Fragen noch ungeklärt sind.

Allgemeine Sprachgeschichte

Leider gibt es nur wenige Arbeiten, die sich mit allgemeinen Tendenzen der Sprachgeschichte beschäftigen. Viele Fragen sind noch unbeantwortet oder werden kontrovers diskutiert.

Probleme

Transparenz

Part of the process of "becoming" a competent Indo-Europeanist has always been recognized as coming to grasp "intuitively" concepts and types of changes in language so as to be able to pick and choose between alternative explanations for the history and development of specific features of the reconstructed language and its offspring.
Schwink (1994)

Anwendbarkeit

- 6909 Sprachen (Ethnologue)
- 128 Sprachfamilien (Ethnologue)
- 47734281 Sprachpaare, die verglichen werden können!

Adäquatheit

Einmal ist keinmal, zweimal ist immer!
Ein Mathematiker über den Umgang der Indogermanisten mit Wahrscheinlichkeiten

Zusammenfassung

- keine verbindliche und transparente Methodik
- größtenteils “literarische Form” der Wissensrepräsentation
- mangelnde Validität der Ergebnisse

Beispiele

Rekonstruktion

Chinesisch *hùi* 翽 “surren” < Altchinesisch *$\mathrm{q}^{\mathrm{wh}\text{ʕ}}$at-s (Baxter und Sagart 2011)

Etymologie

Frucht. Sf std. (9. Jh.), mhd. *vruht*, ahd. *fruht*, as. *fruht*. Entlehnt aus l. *frūctus* m. gleicher Bedeutung (zu l. *fruī* “genieße”). Das deutsche Wort ist Femininum geworden im Anschluß an die *ti*- Abstrakta wie **Flucht²** usw. Adjektive: **fruchtig**, **fruchtbar**; Verb: **(be-)fruchten**. Ebenso nndl. *vrucht*, ne. *fruit*, nfrz. *fruit*, nschw. *frukt*, nnorw. *frukt*; **frugal**.
(Kluge und Seebold 2002)

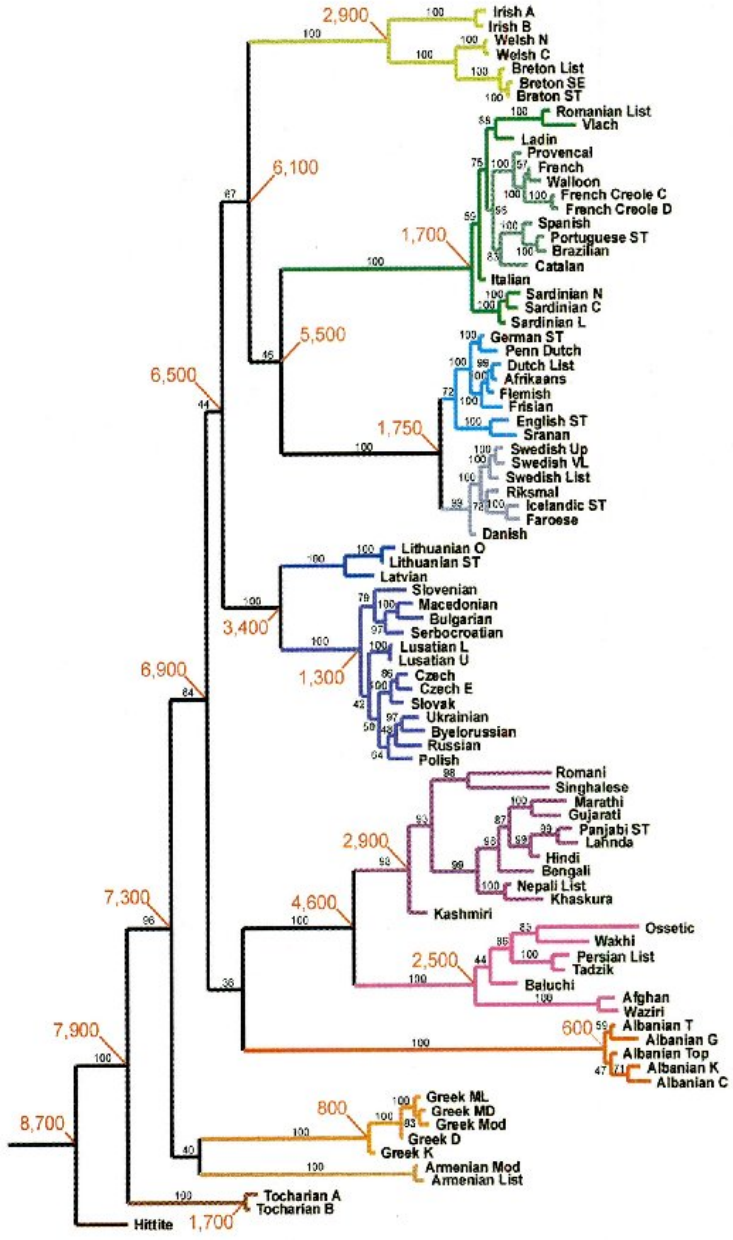
Die quantitative Wende

Language-tree divergence times support the Anatolian theory of Indo-European origin

Russell D. Gray & Quentin D. Atkinson

Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1020, New Zealand

Languages, like genes, provide vital clues about human history[1,2]. The origin of the Indo-European language family is "the most intensively studied, yet still most recalcitrant, problem of historical linguistics"[3]. Numerous genetic studies of Indo-European origins have also produced inconclusive results[4,5,6]. Here we analyse linguistic data using computational methods derived from evolutionary biology. We test two theories of Indo-European origin: the 'Kurgan expansion' and the 'Anatolian farming' hypotheses. The Kurgan theory centres on possible archaeological evidence for an expansion into Europe and the Near East by Kurgan horsemen beginning in the sixth millennium BP[7,8]. In contrast, the Anatolian theory claims that Indo-European languages expanded with the spread of agriculture from Anatolia around 8,000–9,500 years BP[9]. In striking agreement with the Anatolian hypothesis, our analysis of a matrix of 87 languages with 2,449 lexical items produced an estimated age range for the initial Indo-European divergence of between 7,800 and 9,800 years BP. These results were robust to changes in coding procedures, calibration points, rooting of the trees and priors in the bayesian analysis.



Charakteristik

Ursprung

- “Indo-European and computational cladistics” (Ringe, Warnow and Taylor 2002)
- “Language-tree divergence times support the Anatolian theory of Indo-European origin” (Gray und Atkinson 2003)
- “Language classification by numbers” (McMahon und McMahon 2005)
- “Curious Parallels and Curious Connections: Phylogenetic Thinking in Biology and Historical Linguistics” (Atkinson und Gray 2005)
- “Automated classification of the world’s languages” (Brown et al. 2008)
- “Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction” (Downey et al. 2008)
- “Networks uncover hidden lexical borrowing in Indo-European language evolution” (Nelson-Sathi et al. 2011)
- “A pipeline for computational historical linguistics” (Steiner, Stadler, und Cysouw 2011)

Schwerpunkte und Ziele

Schwerpunkte

- Phylogenetische Rekonstruktion (genetische Klassifikation)
- Automatische Sequenzvergleiche
- Allgemeine Fragen der Sprachentwicklung

Ziele

If we cannot guarantee getting the same results from the same data considered by different linguists, we jeopardize the essential scientific criterion of repeatability. (McMahon und McMahon 2005)

Methoden, Theorien und Modelle

Phylogenetische Rekonstruktion

Es gibt eine Vielzahl unterschiedlichster Algorithmen zur phylogenetischen Rekonstruktion. Gemeinsam haben alle, dass Objekte (Sprachen) auf der Grundlage quantitativer Daten (Distanz- oder Ähnlichkeitswerte, Present-Absent-Matrizzen) geclustert werden.

Cognate-Sets (“Kognatensätze”)

Cognate-Sets sind Gruppen von Wörtern unterschiedlicher Sprache, die etymologisch verwandt (kognat, homolog) sind, also ein gemeinsames Vorgängerwort aufweisen. Cognate-Sets spielen eine wichtige Rolle in fast allen neuen quantitativen Ansätzen.

Sequenzalinierung

In einer Alinierungsanalyse werden Sequenzen in einer Matrix dergestalt angeordnet, dass einander entsprechende Segmente in der gleichen Spalte auftauchen, während Null-Entsprechungen durch spezifische Gapsymbole dargestellt werden.

Errungenschaften

Neue Perspektiven

- äußere Sprachgeschichte rückt in den Mittelpunkt
- Abkehr vom traditionellen “Indo-Eurozentrismus”
- neue Fragen der allgemeinen Sprachgeschichte
- neue Modelle der Sprachgeschichte

Neue Ansätze

- empirische Daten rücken in den Mittelpunkt
- stochastische Herangehensweise
- Datenbanken anstelle von Fließtextsammlungen
- Automatisierung der “informellen” Methoden

Probleme

Datenprobleme (Geisler und List, im Druck)

Lexikostatistik (Grundannahmen)

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.

Datenprobleme (Geisler und List, im Druck)

Lexikostatistik (Arbeitsschritte)

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

Datenprobleme (Geisler und List, im Druck)

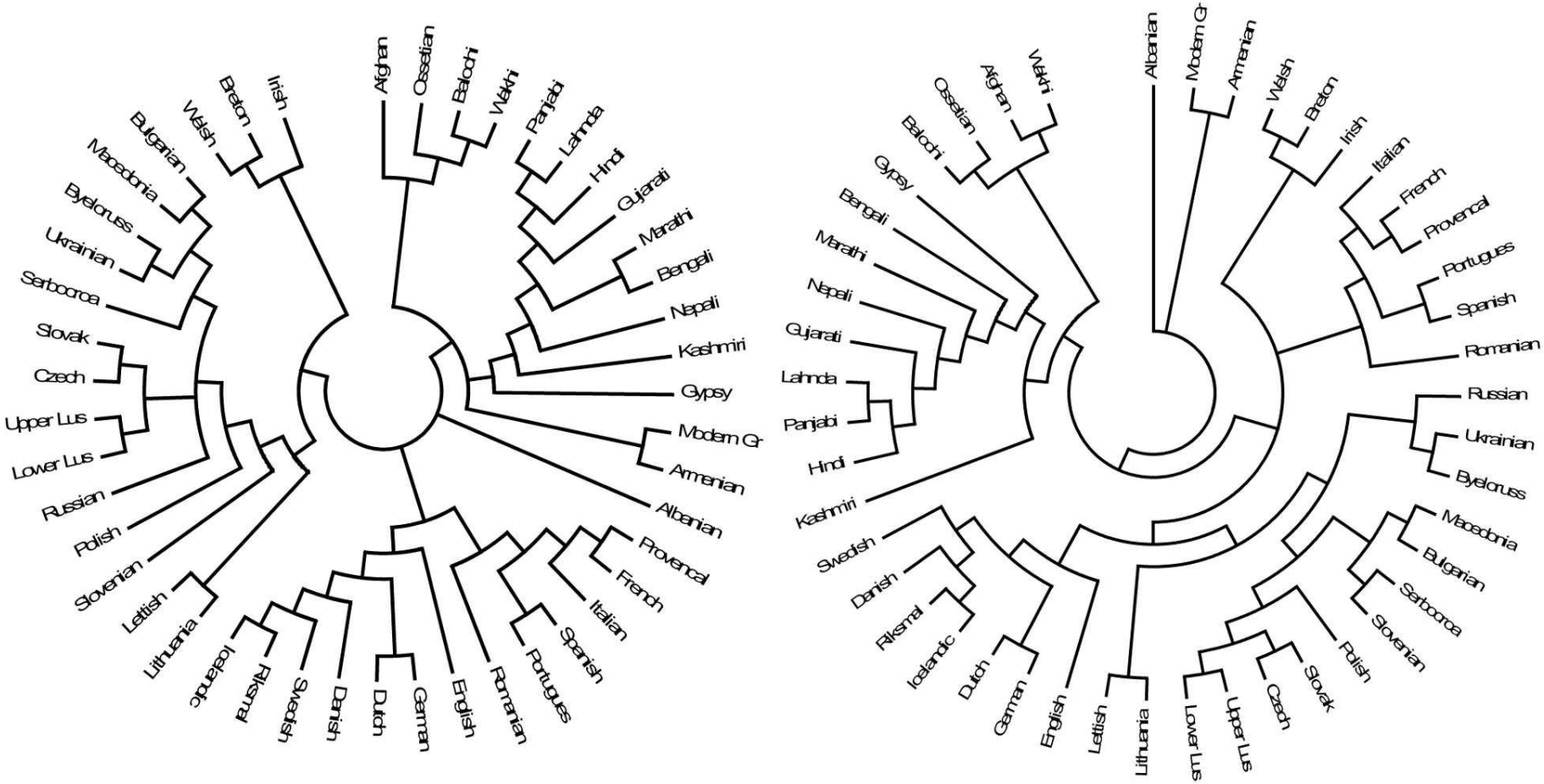
Vergleich zweier Datensätze (Dyen et al. 1997 und Tower of Babel)

Datenbank	Anz. an Spr.	Anz. an Items
Dyen et al. 1997	95	200
Tower of Babel	98	110
Schnittmenge	46	103

Ergebnisse

- bis zu 10 % Unterschiede in Schritt 2 (item translation)
- viele unentdeckte Entlehnungen
- mehr als 30 % Unterschiede in den Baumtopologien (Split-Differenzen)

Datenprobleme (Geisler und List, im Druck)



Fazit

- Viele quantitative Methoden beruhen auf **qualitativ** erstellten Daten.
- Die Methoden zur Erstellung der neuen Daten sind **uneinheitlich** und **fehleranfällig**.
- Die quantitativen Methoden können diese Fehler **nicht** ausmerzen.

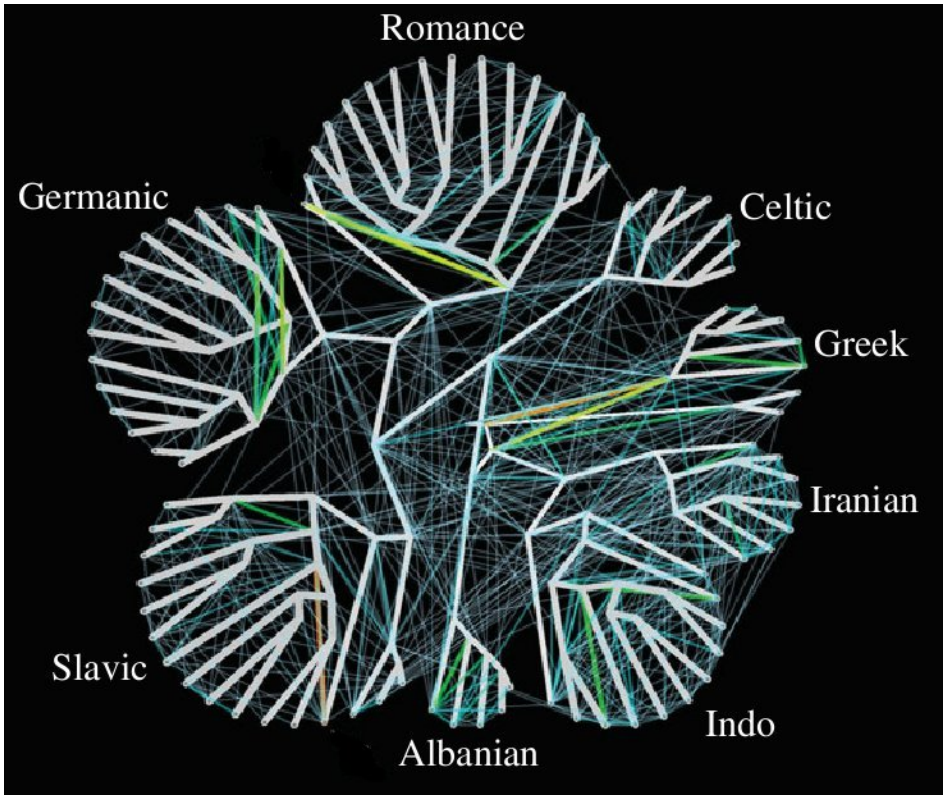
Auf dem Weg zu einer qualitativen Wende?

Taxon Alignment

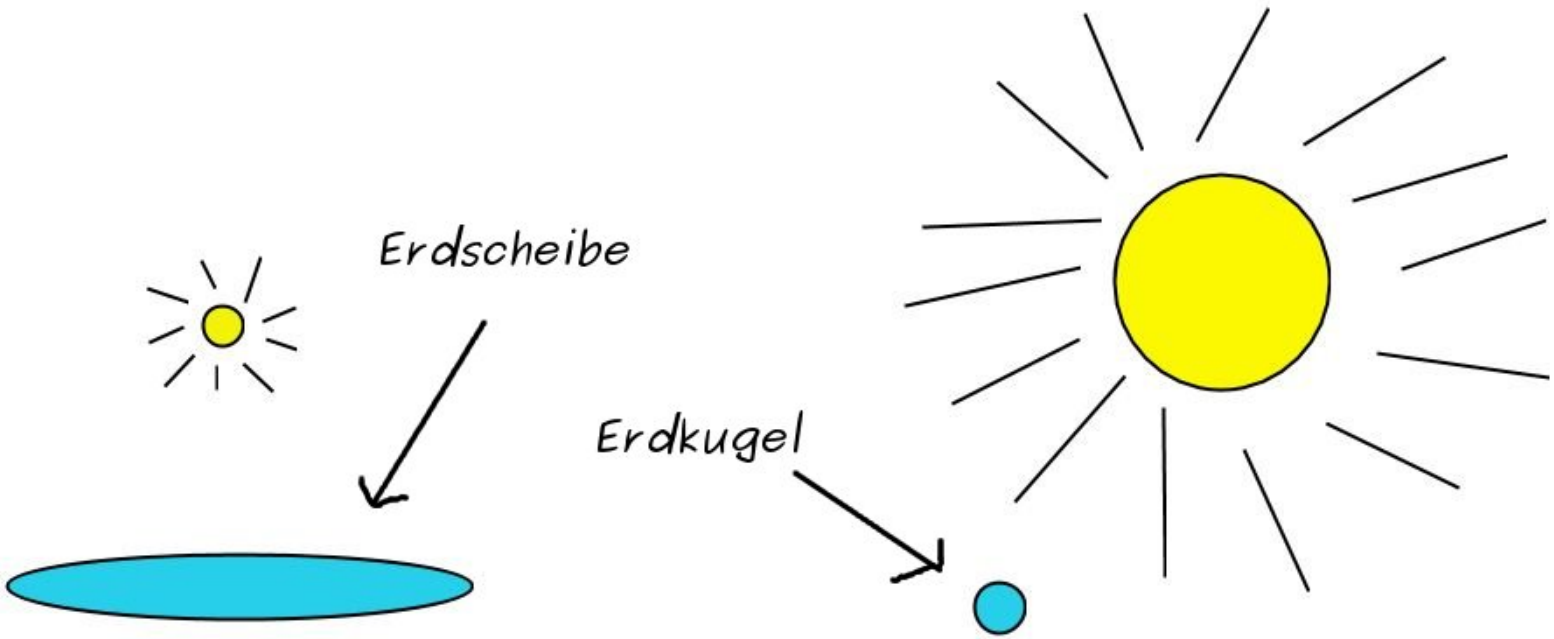
dsh	t͡ʂ	-	o	33	-	-	-
tol	t͡ʂ	-	ɤ	42	-	-	-
ery	t͡ʂ	-	ɤ	44	-	-	-
mzl	ts	-	u	44	-	-	-
xgy	ts	-	ɛ	44	-	-	-
jnc	ts	-	ɤ	44	-	-	-
ggx	ts	w	a	24	-	-	-
lnp	ts	w	a	55	ts	ɯ	33
heq	ts	w	ɔ	44	-	-	-
enq	tʂ	-	o	55	-	-	-
lbz	tʂ	-	o	55	ɖ	ɯ	31
ega	tʂ	-	u	44	-	-	-
jnm	tʂ	-	u	55	-	-	-

Basic Concept: *know (V)* (ID: 45)

CogID	Language	Entry	Aligned Entry
70	German	vɪsən	v ɪ s ə n
70	Danish	viːðə	v iː ð ə -
70	Icelandic	vɪːta	v ɪː t a -
70	Dutch	ʋeːtə	ʋ eː t ə -
70	Norwegian	ʋiːtə	ʋ iː t ə -
70	Swedish	veta	v e t a -
71	English	nəʊ	- - n əʊ -
71	Swedish	ɕɛnːa	ɕ ɛ nː a -
71	German	kɛnən	k ɛ n ə n
71	Danish	kɛnə	k ɛ n ə -
71	Dutch	kɛnə	k ɛ n ə -
71	Norwegian	çenə	ç e n ə -



Paradigmenwechsel



Bioparallelen

Parallelen nach Pagel (2009)

Aspekt	**Spezies**	**Sprachen**
Einheit der Vererbung	Gen	Wort
Replikation	asexuelle und sexuelle Reproduktion	Lernen
Speziation	Kladogenese	Sprachspaltung
Wandelkräfte	natürliche Selektion und genetischer Drift	soziale Selektion und Trends
Differenzierung	baumartig	baumartig

Bioparallelen

Unterschiede

Aspekt	**Spezies**	**Sprachen**
Domäne	Poppers Welt I	Poppers Welt III
Beziehung zw. Form und Funktion	mechanisch	**arbiträr**
Ursprung	Monogenese	**unklar**
Ähnlichkeit zw. Sequenzen	universell (spezies-unabhängig)	**sprachspezifisch**
Differenzierung	baumartig	**netzwerkartig**

Diese Unterschiede werden in den meisten der bisher veröffentlichten neuen Methoden ignoriert.

Terminologie

Homologie und Kognazität

Definition	**Biologie**	**Linguistik**
Gemeinsamer Vorgänger	Homologie	-
Gemeinsamer Direkter Vorgänger	Orthologie	Kognazität
Indirekter gemeinsamer Vorgänger	Paralogie	Kognazität
Lateraler Transfer	Xenologie	Entlehnung

Im Gegensatz zur Biologie hat es die Linguistik bisher versäumt, ein terminologisches Gerüst für historische Zeichenrelationen aufzubauen.

Beispiele

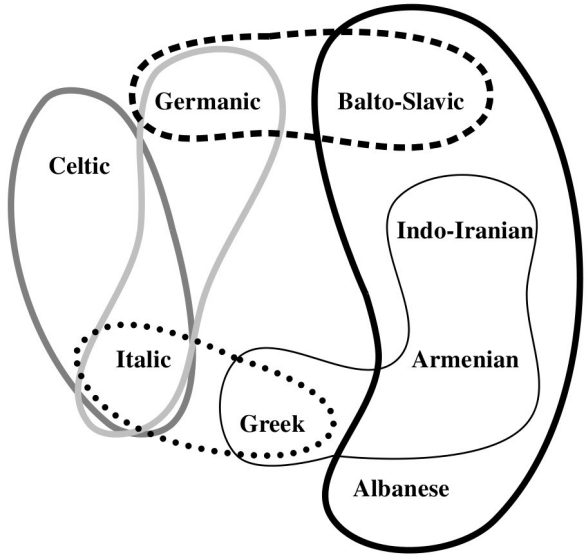

facebook
E-Mail oder Telefon
Passwort
Anmelden
Angemeldet bleiben
Passwort vergessen?
Registrieren
Facebook ermöglicht es dir, mit den Menschen in deinem Leben in Verbindung zu treten und Inhalte mit diesen zu teilen.
MUSTER
Max Mustermann
Gefällt mir
Seite erstellen
Interesse
Beschreibung
Von Wikipedia, die kostenlose Enzyklopädie
Max Mustermann in Österreich und Deutschland, Erika Mustermann in Deutschland, Felix Muster oder Maria Bernasconi in der Schweiz, sind fiktive Personen, die als Beispielnamen – unter anderem in Formularen, Hinweisen und Datenbanken – verbreitet werden. Die Namen werden aber auch als Bezeichnungen für Durchschnittsbürger verwendet.
Erika Mustermann
Als typische Angaben für sie wird als Geburtsdatum der 12. September 1945 in München und nach anderer Quelle der 12. August 1964 erwähnt. Ihr Geburtsname ist Erika Gabler. 1981 war sie wohnhaft in München, Heidestraße 17 (1995 aber „unbekannt verzogen"), 2007 wohnte Frau Mustermann wieder in der Heidestraße 17, allerdings diesmal in Köln, 2008 in Berlin. Beim Geburtsort ist man sich nicht sicher, der Ausweis nennt Berlin, in den verschiedenen Reisepässen taucht neben Berlin auch wieder München auf. Frau Mustermann ist 1,60 m groß und hat grüne Augen.
Info
Wikipedia
594
gefällt das
9
sprechen darüber
Du möchtest angeben, dass dir diese Seite gefällt?
Um mit Max Mustermann interagieren zu können, musst du dich zunächst bei Facebook registrieren.
Registrieren
Facebook ist kostenlos und jeder kann sich registrieren. Du bist bereits ein Mitglied? Melde dich an.

Phylogenetische Netzwerke (Nelson-Sathi et al. 2011)

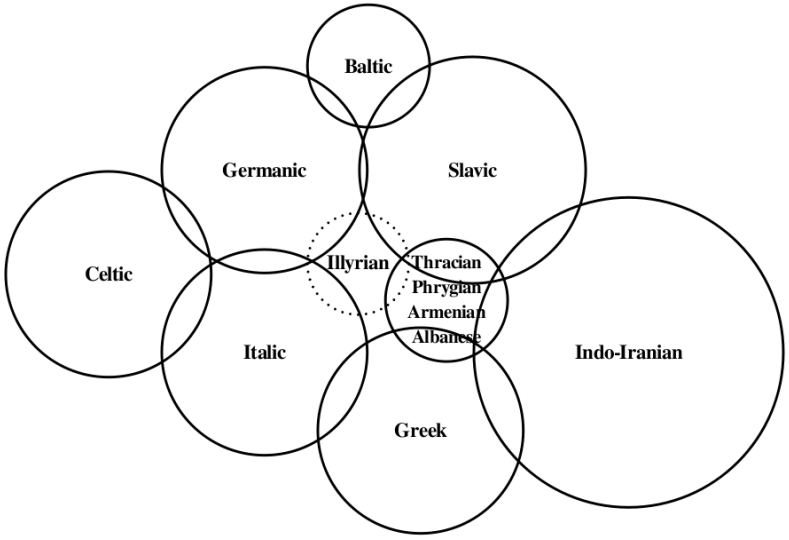
- Das grundlegende Modell zur genetischen Sprachklassifikation ist das Stammbaummodell (Schleicher 1853).
- Dieses genießt jedoch kein volles Vertrauen in der historischen Linguistik und wurde in einer Vielzahl von Arbeiten bereits sehr früh kritisiert (Schuchardt 1870, Schmidt 1872).
- Hauptkritikpunkte betreffen die Praktikabilität, die Plausibilität und die Adäquatheit des Modells.
- Ein Großteil der Kritik bezieht sich auf die Praktikabilität.
- Alternative Modelle wurden unter dem Schlagwort “Wellentheorie” (Schmidt 1872) postuliert, jedoch konnte keiner dieser Ansätze sich durchsetzen.

Phylogenetische Netzwerke (Nelson-Sathi et al. 2011)

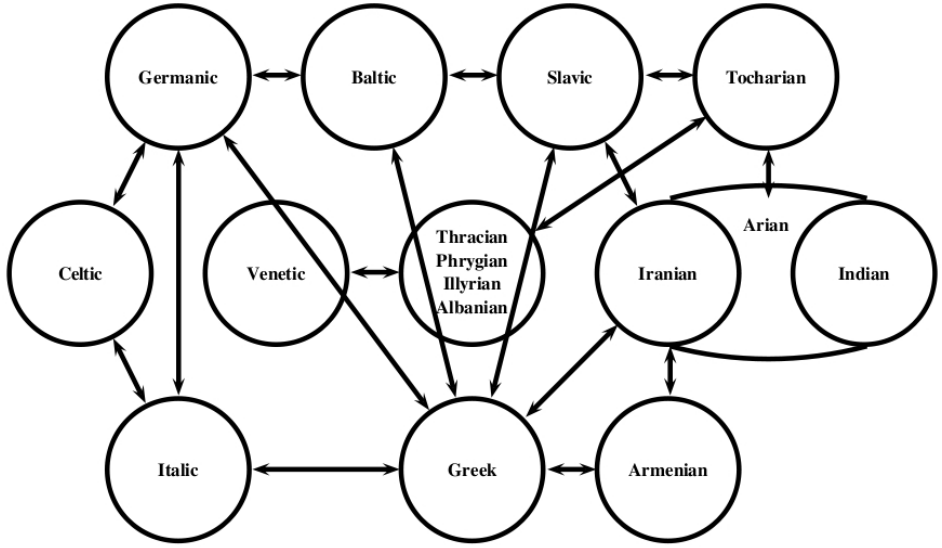
Meillet (1908)



Bloomfield (1933)



Hirt (1905)



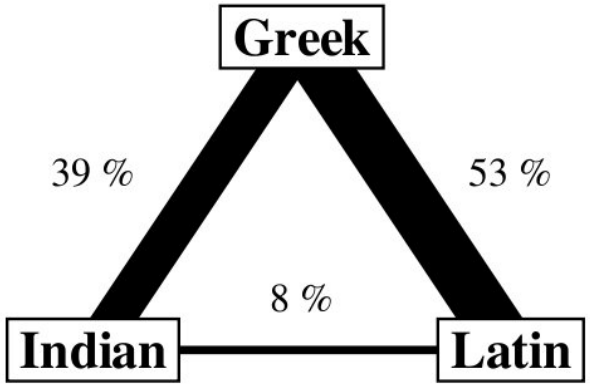
Bonfante (1931)

Phylogenetische Netzwerke (Nelson-Sathi et al. 2011)

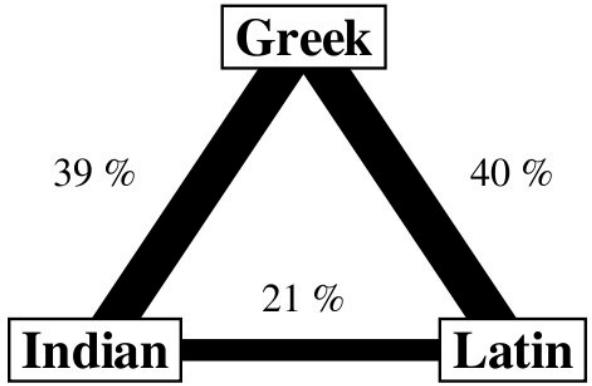
Kritik an der Praktikabilität

Viele Forscher propagierten die Welle als Alternative zum Baum, weil sie die Praktikabilität der Bäume bezweifelten (Schmidt 1872, Bonfante 1933). Streng genommen reicht derartige Kritik jedoch nicht aus, da Praktikabilität durch verbesserte Methoden gesteigert werden kann.

Schmidt (1872)



Nicholaev (2007)



Phylogenetische Netzwerke (Nelson-Sathi et al. 2011)

Phylogenetische Netzwerke

Angesichts der großen Bedeutung lateraler Beziehungen im Verlaufe der Sprachgeschichte, scheint das Baummodell nicht angemessen zu sein, Sprachgeschichte realistisch abzubilden. Phylogenetische Netzwerke sind eine realistischere Alternative, insofern als sie sowohl laterale als auch vertikale Beziehungen zwischen Taxa darstellen können.

> *Wir verbinden die Äste und Zweige des Stammbaums durch zahllose horizontale Linien, und er hört auf ein Stammbaum zu sein.* (Schuchardt 1870)

Phylogenetische Netzwerke (Nelson-Sathi et al. 2011)

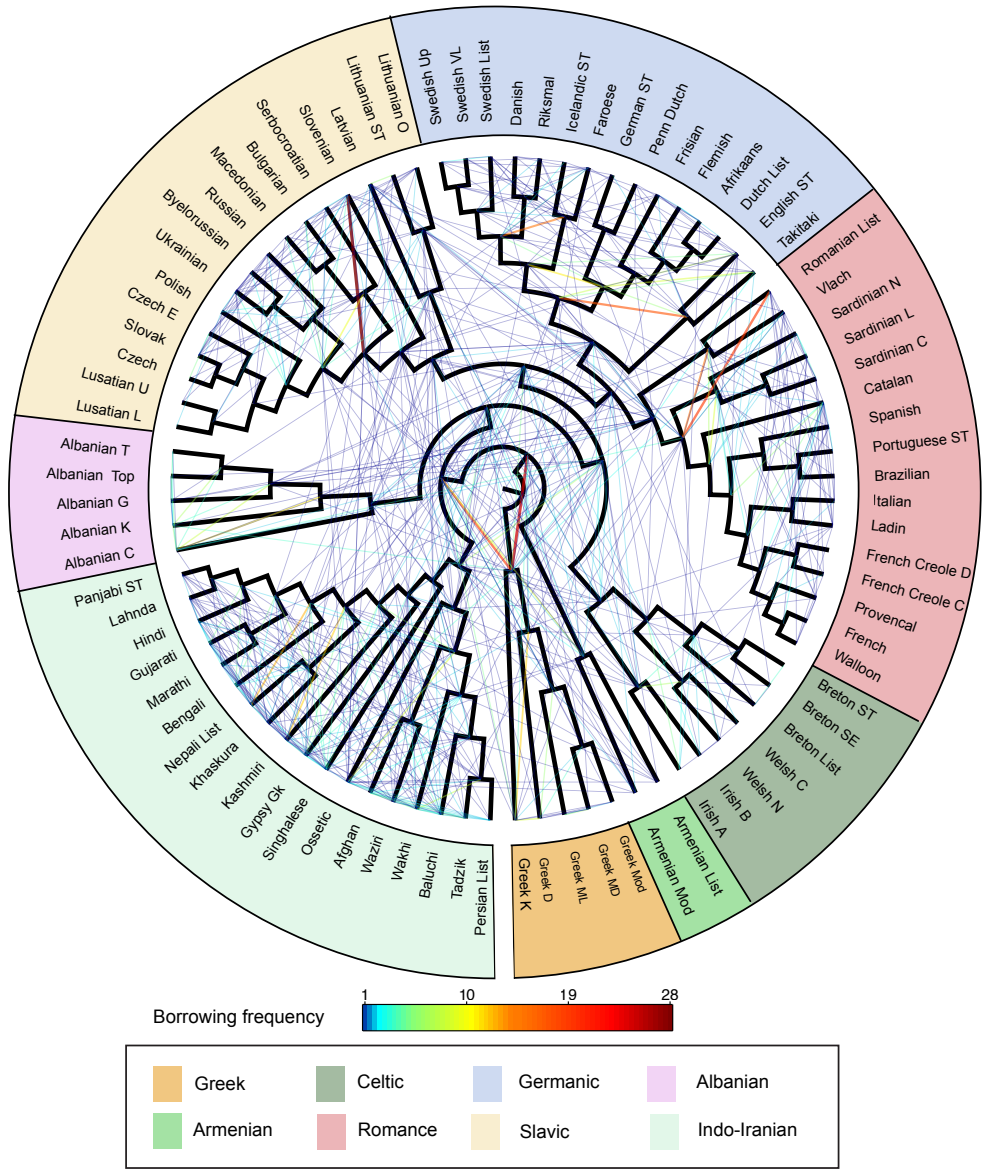


Fig. Minimal Lateral Network (MLN) of 84 Indo-European languages.

Lautklassenbasierte Alinierung (List, im Druck)

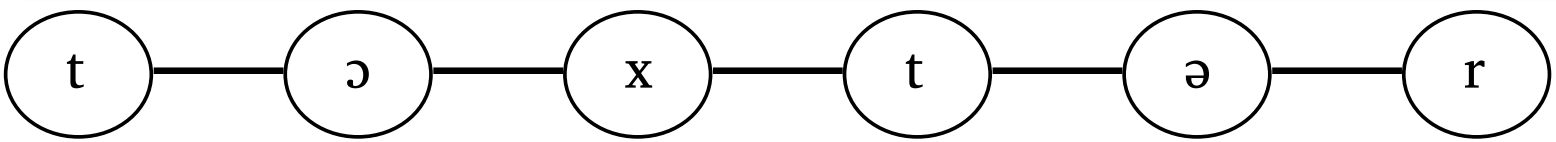
Alinierung

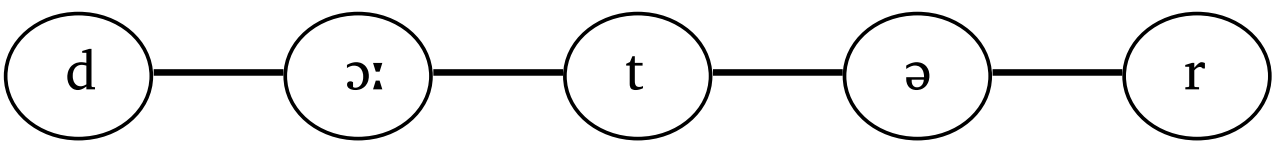
In einer Alinierungsanalyse werden Sequenzen in einer Matrix dergestalt angeordnet, dass einander entsprechende Segmente in der gleichen Spalte auftauchen, während Null-Entsprechungen durch spezifische Gapsymbole dargestellt werden.

Lautklassenbasierte Alinierung (List, im Druck)

Alinierung

In einer Alinierungsanalyse werden Sequenzen in einer Matrix dergestalt angeordnet, dass einander entsprechende Segmente in der gleichen Spalte auftauchen, während Null-Entsprechungen durch spezifische Gapsymbole dargestellt werden.

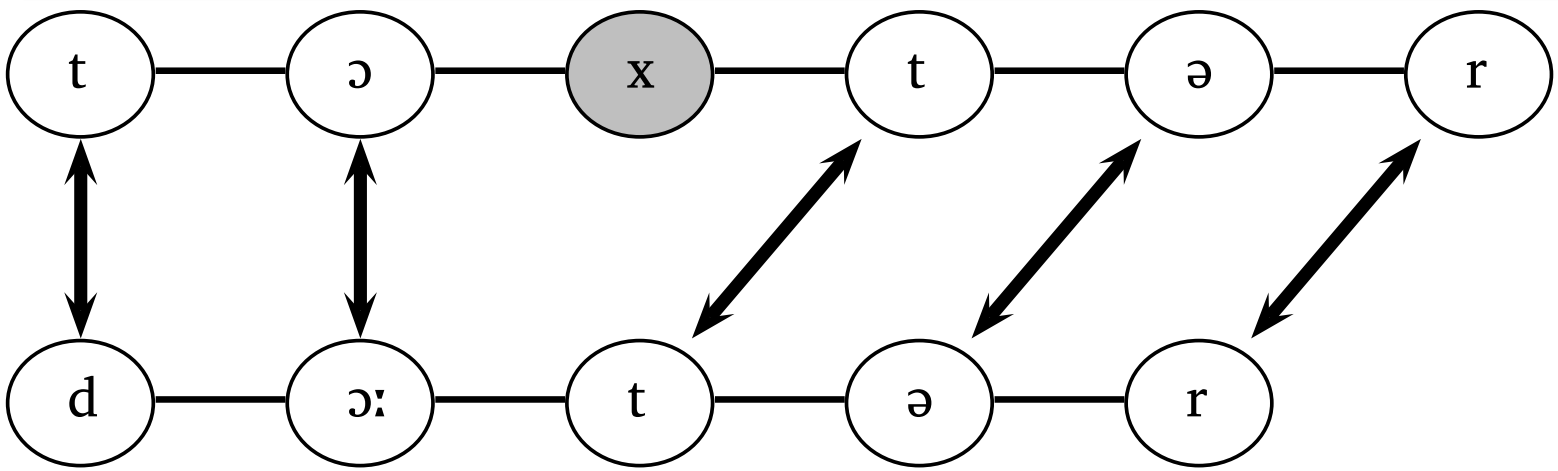




Lautklassenbasierte Alinierung (List, im Druck)

Alinierung

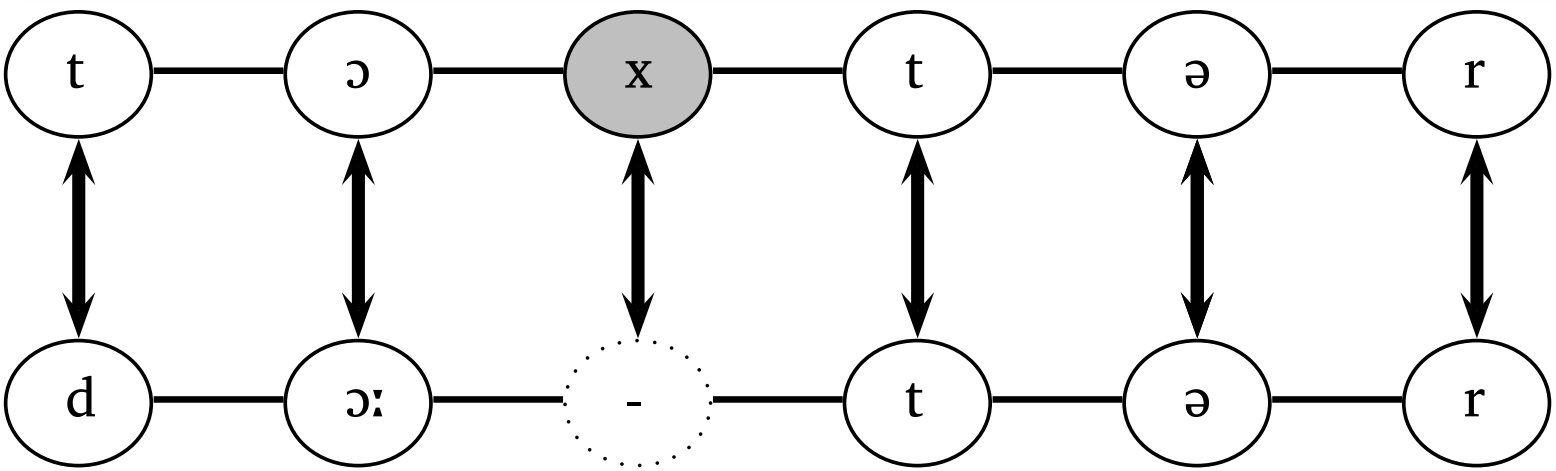
In einer Alinierungsanalyse werden Sequenzen in einer Matrix dergestalt angeordnet, dass einander entsprechende Segmente in der gleichen Spalte auftauchen, während Null-Entsprechungen durch spezifische Gapsymbole dargestellt werden.



Lautklassenbasierte Alinierung (List, im Druck)

Alinierung

In einer Alinierungsanalyse werden Sequenzen in einer Matrix dergestalt angeordnet, dass einander entsprechende Segmente in der gleichen Spalte auftauchen, während Null-Entsprechungen durch spezifische Gapsymbole dargestellt werden.



Lautklassenbasierte Alinierung (List, im Druck)

Lautklassen

Laute, die häufig in Korrespondenzbeziehung in genetisch verwandten Sprachen stehen, können in Klassen zusammengefasst werden. Es wird dabei angenommen, dass "phonetic correspondences inside a 'type' are more regular than those between different 'types'" (Dolgopolsky 1986: 35).

Lautklassenbasierte Alinierung (List, im Druck)

Lautklassen

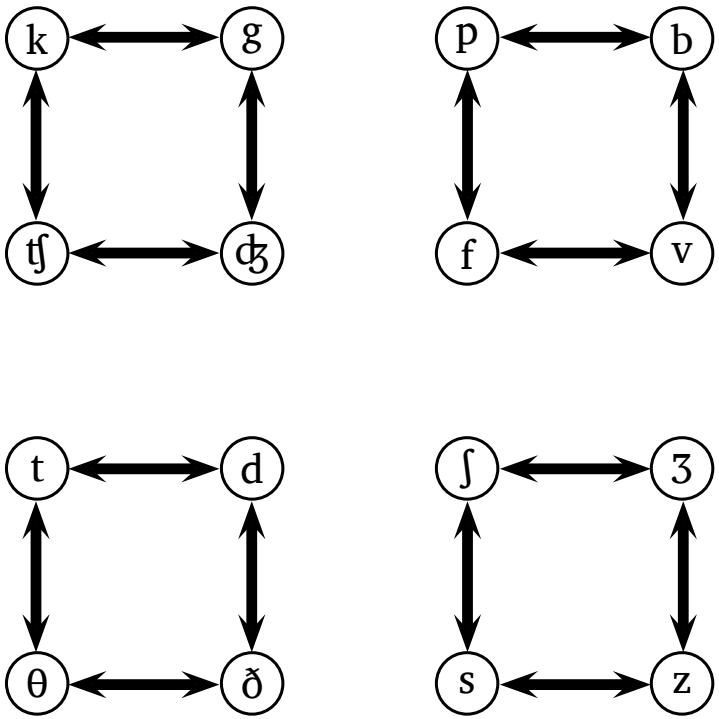
Laute, die häufig in Korrespondenzbeziehung in genetisch verwandten Sprachen stehen, können in Klassen zusammengefasst werden. Es wird dabei angenommen, dass “phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’” (Dolgopolsky 1986: 35).

k	g	p	b
ʧ	ʤ	f	v
t	d	ʃ	ʒ
θ	ð	s	z

Lautklassenbasierte Alinierung (List, im Druck)

Lautklassen

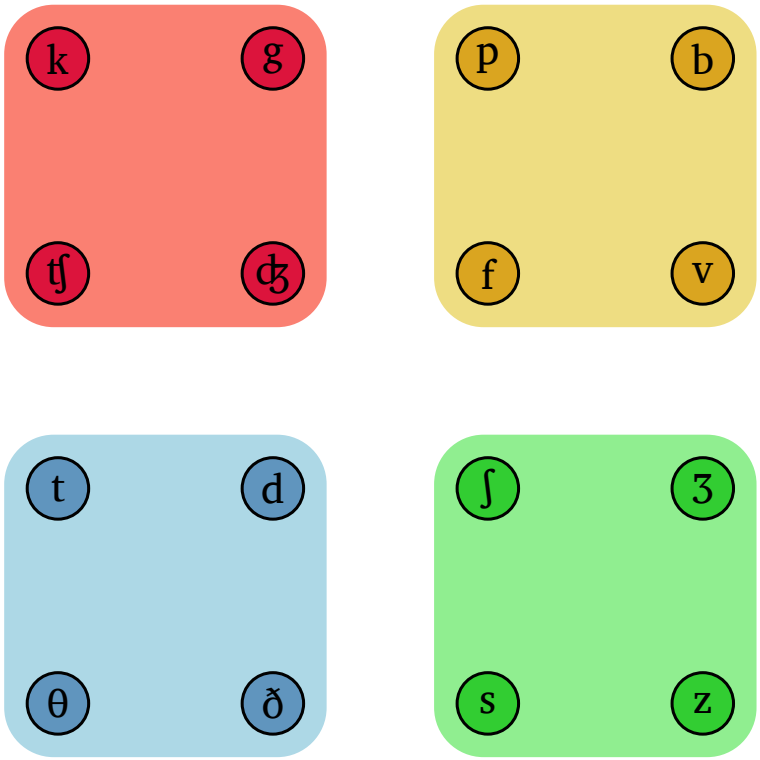
Laute, die häufig in Korrespondenzbeziehung in genetisch verwandten Sprachen stehen, können in Klassen zusammengefasst werden. Es wird dabei angenommen, dass “phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’” (Dolgopolsky 1986: 35).



Lautklassenbasierte Alinierung (List, im Druck)

Lautklassen

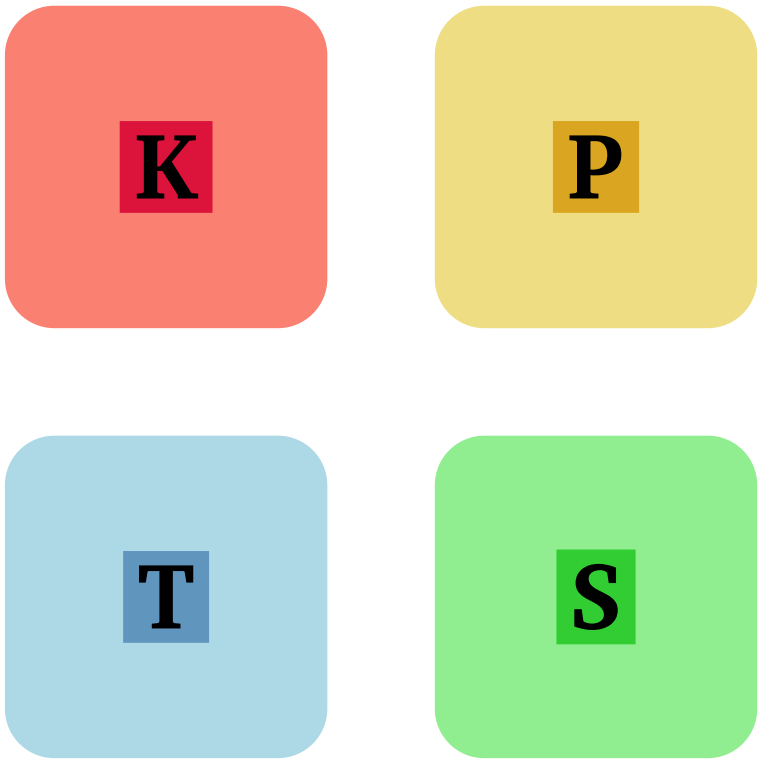
Laute, die häufig in Korrespondenzbeziehung in genetisch verwandten Sprachen stehen, können in Klassen zusammengefasst werden. Es wird dabei angenommen, dass "phonetic correspondences inside a 'type' are more regular than those between different 'types'" (Dolgopolsky 1986: 35).



Lautklassenbasierte Alinierung (List, im Druck)

Lautklassen

Laute, die häufig in Korrespondenzbeziehung in genetisch verwandten Sprachen stehen, können in Klassen zusammengefasst werden. Es wird dabei angenommen, dass "phonetic correspondences inside a 'type' are more regular than those between different 'types'" (Dolgopolsky 1986: 35).



Lautklassenbasierte Alinierung (List, im Druck)

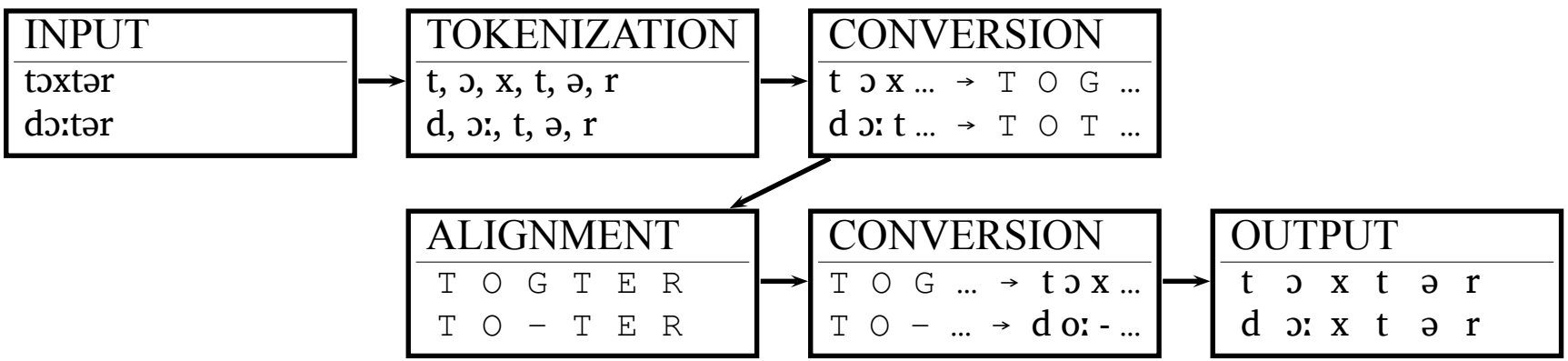
Sound-Class-Based Phonetic Alignment (SCA)

Lautklassen und Alinierungsanalysen können einfach kombiniert werden, indem Lautsequenzen intern als Lautklassen repräsentiert, und diese Lautklassen dann mit Hilfe traditioneller Algorithmen aliniert werden.

Lautklassenbasierte Alinierung (List, im Druck)

Sound-Class-Based Phonetic Alignment (SCA)

Lautklassen und Alinierungsanalysen können einfach kombiniert werden, indem Lautsequenzen intern als Lautklassen repräsentiert, und diese Lautklassen dann mit Hilfe traditioneller Algorithmen aliniert werden.



Lautklassenbasierte Alinierung (List, im Druck)

- Die neueste Version SCA-Methode erreicht eine Akkurazität von über 90 % für multiple Alinierungsanalysen.
- Die SCA-Methode kann für alle sprachlichen Daten angewendet werden (inklusive Tonsprachen), solange diese in phonetischer Transkription vorliegen.
- Die SCA-Methode erlaubt es, über spezifische Visualisierungstechniken, die von der Evolutionsbiologie inspiriert wurden, einen neuen Blick auf Wortähnlichkeiten zu werfen.

Lautklassenbasierte Alinierung (List, im Druck)

Taxon	Alignment									
Xiangtan	-	i	-	24	d	əɯ	12	-	-	-
Qingdao	-	i	-	42	tʰ	ou	-	-	-	-
Xi_an	-	ə	r	21	tʰ	ou	-	-	-	-
Wuhan	-	ɯ	-	213	tʰ	əu	-	-	-	-
Guangzhou	j	i	t	2	tʰ	ɐu	21	-	-	-
Nanning	j	i	t	22	tʰ	ɐu	21	-	-	-
Xianggang	j	ɐ	t	2	tʰ	ɐu	21	-	-	-
Nanning	j	ɐ	t	22	tʰ	ɐu	21	-	-	-
Taibei	l	i	t	44	tʰ	au	24	-	-	-
Xiamen	l	i	t	5	tʰ	au	-	-	-	-
Jian_ou	m	i	-	33	tʰ	e	33	-	-	-
Shexian	n	i	-	22	tʰ	iu	44	-	-	-
Fuzhou	n	i	ʔ	5	tʰ	au	53	-	-	-
Shantou	z	i	k	5	-	-	-	-	-	-
Shantou	z	i	k	5	tʰ	au	55	-	-	-
Haikou	z	i	t	3	-	-	-	-	-	-
Haikou	z	i	t	3	h	au	31	-	-	-
Taoyuan	ŋ	i	t	22	tʰ	eu	11	-	-	-
Meixian	ŋ	i	t	1	tʰ	eu	11	-	-	-
Wenzhou	nʱ	i	-	213	d	ɤu	-	-	-	-
Wenzhou	nʱ	i	-	213	d	ɤu	31	v	ai	213
Nanchang	nʱ	i	ʔ	5	tʰ	u	02	-	-	-
Tunxi	nʱ	ie	-	11	tʰ	iu	44	-	-	-
Suzhou	nʱ	iə	ʔ	3	d	ʏ	13	-	-	-
Shanghai	nʱ	ɪ	ʔ	1	d	ɤ	13	-	-	-
Kunming	ʐ	ə	-	31	tʰ	au	31	-	-	-
Hefei	ʐ	ə	ʔ	5	tʰ	ɯ	-	-	-	-
Xining	ʐ	ɛ	-	44	tʰ	ɯ	24	-	ɛ	24
Jinan	ʐ	ɻ	-	21	tʰ	ou	-	-	-	-
Changsha	ʐ	ɻ	-	24	t	əu	-	-	-	-
Zhengzhou	ʐ	ɻ	-	24	tʰ	ou	-	-	-	-
Lanzhou	ʐ	ɻ	-	13	tʰ	ou	13	-	-	-
Yinchuan	ʐ	ɻ	-	13	tʰ	əu	-	-	-	-
Haerbin	ʐ	ɻ	-	53	tʰ	ou	-	-	-	-
Beijing	ʐ	ɻ	-	51	tʰ	ou	1	-	-	-
Huhehaote	ʐ	ɻ	-	55	tʰ	əu	31	-	-	-
Pingyao	ʐ	ʌ	ʔ	53	t	əu	13	-	iᴇ	13

Automatische Kognatenerkennung (List 2012)

Die komparative Methode

- Erstelle eine Liste möglicher Kognaten.
- Extrahiere eine Liste möglicher Lautkorrespondenzen aus der Kognatenliste.
- Modifiziere und verbessere die beiden listen durch
  - Hinzufügen und Entfernen von Kognatensätzen von der Kognatenliste, in Abhängigkeit davon, ob diese kompatibel sind mit der Korrespondenzliste, und
  - Hinzufügen und Entfernen von Korrespondenzen von der Korrespondenzliste, in Abhängigkeit davon, ob diese kompatibel sind mit der Kognatenliste.
- Veröffentliche die Ergebnisse, wenn sie zufriedenstellend sind.

Automatische Kognatenerkennung (List 2012)

Sprachspezifische Ähnlichkeit

- Sequenzähnlichkeit wird auf der Grundlage systematischer Lautkorrespondenzen bestimmt und nicht auf der Grundlage von oberflächlichen Ähnlichkeiten.
- Lass (1997) nennt diese Ähnlichkeit *genotypisch* im Gegensatz zu einer *phänotypischen* Ähnlichkeit.
- Der wichtigste Aspekt der korrespondenzbasierten Ähnlichkeit ist jedoch, dass sie *sprachspezifisch* ist: Genotypische Ähnlichkeit is nie generell definiert, sondern immer in Bezug auf zwei Sprachsysteme, die miteinander verglichen werden.

Automatische Kognatenerkennung (List 2012)

Sprachspezifische Ähnlichkeit

- Sequenzähnlichkeit wird auf der Grundlage systematischer Lautkorrespondenzen bestimmt und nicht auf der Grundlage von oberflächlichen Ähnlichkeiten.
- Lass (1997) nennt diese Ähnlichkeit *genotypisch* im Gegensatz zu einer *phänotypischen* Ähnlichkeit.
- Der wichtigste Aspekt der korrespondenzbasierten Ähnlichkeit ist jedoch, dass sie *sprachspezifisch* ist: Genotypische Ähnlichkeit is nie generell definiert, sondern immer in Bezug auf zwei Sprachsysteme, die miteinander verglichen werden.

Meaning	German	Dutch	English
“tooth”	*Zahn* [ts aːn]	*tand* [t ɑnt]	*tooth* [t ʊːθ]
“ten”	*zehn* [ts eːn]	*tien* [t iːn]	*ten* [t ɛn]
“tongue”	*Zunge* [ts ʊŋə]	*tong* [t ɔŋ]	*tongue* [t ʌŋ]

Automatische Kognatenerkennung (List 2012)

Sprachspezifische Ähnlichkeit

- Sequenzähnlichkeit wird auf der Grundlage systematischer Lautkorrespondenzen bestimmt und nicht auf der Grundlage von oberflächlichen Ähnlichkeiten.
- Lass (1997) nennt diese Ähnlichkeit *genotypisch* im Gegensatz zu einer *phänotypischen* Ähnlichkeit.
- Der wichtigste Aspekt der korrespondenzbasierten Ähnlichkeit ist jedoch, dass sie *sprachspezifisch* ist: Genotypische Ähnlichkeit is nie generell definiert, sondern immer in Bezug auf zwei Sprachsysteme, die miteinander verglichen werden.

Meaning	Shanghai	Beijing	Guangzhou
“nine”	[tɕ iɤ^{35}]	Beijing [tɕ iou^{214}]	[k ɐu^{35}]
“today”	[tɕ iŋ^{55}tsɔ^{21}]	Beijing [tɕ iɚ^{55}]	[k ɐm^{53}jɐt^{2}]
“rooster”	[koŋ^{55} tɕ i^{21}]	Beijing[kuŋ^{55} tɕ i^{55}]	[k ɐi^{55}koŋ^{55}]

Automatische Kognatenerkennung (List 2012)

LexStat

LexStat ist eine Methode zur automatischen Kognatenerkennung in mehrsprachigen Wortlisten. LexStat basiert auf lautklassenbasierter Sequenzalinierung, mit deren Hilfe sprachspezifische Lautähnlichkeiten (ähnlich den regulären Lautkorrespondenzen) identifiziert werden. Basierend auf diesen sprachspezifischen Ähnlichkeitsmaßen werden Wörter in Kognatensätze geclustert. Die Methode erreicht für kleine Datensätze eine Akkurazität von 85 % und ist damit viel zuverlässiger als simple Alinierungsmethoden (76 %). An größeren Datensätzen konnte die Method noch nicht getestet werden, weil diese erst noch erstellt werden müssen. Es ist jedoch davon auszugehen, dass die Akkurazität bei größeren Datensätzen weiter steigt. Wie auch die SCA-Methode ist LexStat universell auf alle Sprachen anwendbar, für die phonetische Daten (IPA) vorliegen.

Automatische Kognatenerkennung (List 2012)

ID	Items	German	English	Swedish
1	hand	hant	hænd	hand
2	woman	fraʊ	wʊmən	kvina
3	know	kɛnən	nəʊ	çɛna
3	know	vɪsən	-	ve:ta
…	…	…	…	…

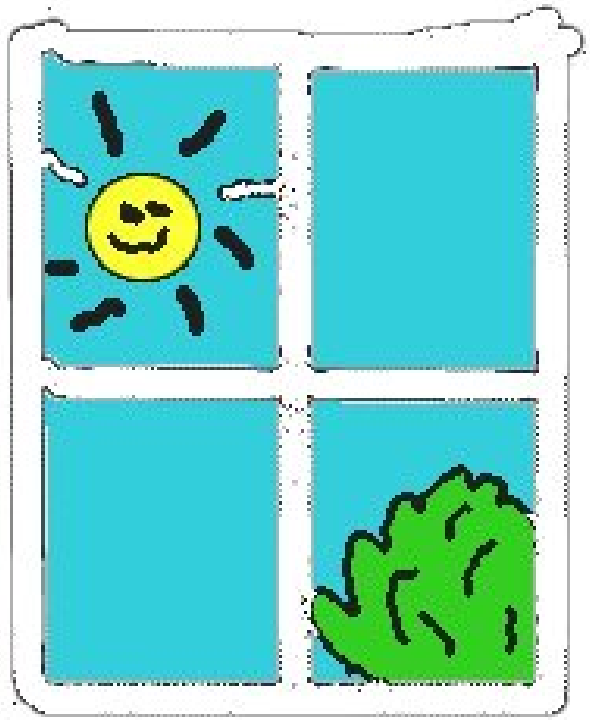
Automatische Kognatenerkennung (List 2012)

ID	Items	German	COG	English	COG	Swedish	COG
1	hand	hant	1	hænd	1	hand	1
2	woman	fraʊ	2	wʊmən	3	kvina	4
3	know	kɛnən	5	nəʊ	5	çɛna	5
3	know	vɪsən	6	-	0	ve:ta	6
…	…	…	…	…	…	…	…

Automatische Kognatenerkennung (List 2012)

Basic Concept: *belly* (ID: 4)			
CogID	**Language**	**Entry**	**Aligned Entry**
6	Danish	ɔnʌliwˀ	--
7	German	baux	b au x
7	Dutch	bœyk	b œy k
7	Swedish	buk	b u k
7	Norwegian	bʉ:k	b ʉː k
8	English	bɛlɪ	--
9	Swedish	ma:ge	m aː g e
9	Norwegian	mɑ:gə	m ɑː g ə
9	Danish	mæ:və	m æː v ə
10	Icelandic	kʰvɪ:ðʏr	--

Ausblick



Ausblick

Von den Biologen lernen...

- stochastisch gestützte Hypothesen anstelle von impressionistischen, intuitiven “Wahrheiten”
- maschinenlesbare Datensätze anstelle von Informationsvernichtung in Fließtexten
- rigoroses Testen von Algorithmen
- Festlegen einheitlicher Terminologien und Formate
- entspannter Umgang mit Fehlern in den Methoden

Ausblick

Von den Biologen lernen...

- stochastisch gestützte Hypothesen anstelle von impressionistischen, intuitiven “Wahrheiten”
- maschinenlesbare Datensätze anstelle von Informationsvernichtung in Fließtexten
- rigoroses Testen von Algorithmen
- Festlegen einheitlicher Terminologien und Formate
- entspannter Umgang mit Fehlern in den Methoden

Linguist bleiben...

- Parallelen zwischen Biologie und Linguistik müssen kritisch hinterfragt werden
- offensichtliche Unterschiede zwischen Biologie und Linguistik bedürfen der Entwicklung spezifischer, neuer Methoden


Danke fürs Zuhören!