

Gerhard Jäger (Tübingen)

## Lexikostatistik 2.0\*

### Abstract

In der Mitte des 20. Jahrhunderts gab es diverse Versuche, die Klassifikation von Sprachen mit Hilfe von Wortlisten, die dem Grundvokabular der betreffenden Sprachen entnommen sind, zu automatisieren. Diese Methoden wurden und werden in der historischen Sprachwissenschaft gemeinhin kritisch diskutiert, da sich die erzielten Ergebnisse häufig als fehlerhaft erwiesen.

In den letzten Jahren erleben wir einen neuen Aufschwung lexikostatistischer und glottochronologischer Ansätze. Deren Erfolgsaussichten sind heute wesentlich besser als vor einem halben Jahrhundert, da uns jetzt große Mengen an sprachvergleichenden Daten in elektronischer Form zur Verfügung stehen und die Computerlinguistik und Bioinformatik mächtige Werkzeuge bereitstellt, diese Daten statistisch auszuwerten.

Im vorliegenden Artikel wird eine Fallstudie vorgestellt, die das Potenzial lexikostatistischer Methoden im 21. Jahrhundert illustriert.

### 1. Einleitung

Einer der faszinierendsten Forschungsgegenstände der historischen Sprachwissenschaft ist die Frage, in welchen Verwandtschaftsverhältnissen Sprachen zueinander stehen. Das Erkenntnisideal wäre ein Familienstammbaum aller bekannter Sprachen.

Die traditionelle komparative Methode strebt an, Sprachwandelprozesse so weit wie möglich zu rekonstruieren; die Erstellung eines Sprachstammbaums ergibt sich dabei in gewisser Weise als Nebeneffekt. Diese Vorgehensweise hat sich in den vergangenen zwei Jahrhunderten als sehr erfolgreich erwiesen. Besonders für historisch gut dokumentierte Sprachgruppen wie die indoeuropäischen oder die semitischen Sprachen sind die erreichten Rekonstruktionen wie auch die Kenntnisse über die Verwandtschaftsverhältnisse innerhalb dieser Gruppen sehr detailliert.

Allerdings hat die historisch-komparative Methode einige inhärente Begrenzungen. Nicht nur fehlen für die meisten Sprachen der Welt schriftliche Überlieferungen, die zur Rekonstruktion herangezogen werden können. Die Zeittiefe möglicher Rekonstruktionen ist vermutlich auf maximal zehn Jahrtausende begrenzt (wobei es höchst kontrovers ist, ob diese Schallmauer deutlich näher oder vielleicht sogar etwas ferner liegt). Nicht zuletzt ist historisch-komparative Rekonstruktion sehr zeitaufwändig und setzt viel Erfah-

---

\* Diese Forschung wurde im Rahmen des ERC Advanced Grant 324246 *Language Evolution: The Empirical Turn* (EVOLAEMP) durchgeführt.

rung und Expertise voraus. Auch aus diesem Grund ist es unrealistisch zu erwarten, dass unser Wissen über die Geschichte etwa der altamerikanischen Sprachen oder der Sprachen Papua-Neuguineas jemals ähnlich detailliert sein wird wie z.B. das über die indoeuropäischen Sprachen.

Es ist daher lohnenswert zu untersuchen, ob Verwandtschaftsbeziehungen zwischen Sprachen auch dann erhellt werden können, wenn eine vollständige Rekonstruktion der historischen Prozesse nicht möglich ist. In den vergangenen Jahrzehnten hat es immer wieder Versuche gegeben, hier Fortschritte zu machen.<sup>1</sup> Die Lexikostatistik war ein derartiger Versuch, der in der Mitte des zwanzigsten Jahrhunderts von dem amerikanischen Linguisten Morris Swadesh entwickelt wurde. Sie ist seither – möglicherweise zu Unrecht – in Misskredit geraten und wird in Einführungsdarstellungen gerne als Irrweg dargestellt. Allerdings sind wir heute in einer wesentlich komfortableren Position als Swadesh zu seiner Zeit, da sich sowohl der Zugang zu großen Mengen an sprachvergleichenden Daten als auch die Techniken und die Hardware für numerische Auswertungen dieser Daten auf gerade atemberaubende Weise verbessert haben. In diesem Aufsatz möchte ich darlegen, dass Lexikostatistik mit den Mitteln des Jahres 2013 ein durchaus ernstzunehmendes Forschungsprogramm ist.

## 2. Lexikostatistik nach Swadesh

Die von Swadesh vorgeschlagene Vorgehensweise besteht aus vier Schritten (siehe z.B. Swadesh 1971):

- 1) Erstellung einer Liste von Konzepten, die in (nahezu) allen Sprachen lexikalisiert werden und deren Lexikalisierungen selten zwischen Sprachen entlehnt werden.
- 2) Sammlung von Wörtlisten, also der Lexikalisierungen dieser Konzeptliste, für die zu untersuchenden Sprachen.
- 3) Bestimmung, welche synonymen Ausdrücke aus verschiedenen Sprachen (innerhalb dieser Wörtlisten) kognat<sup>2</sup> zueinander sind.
- 4) Für jedes Sprachpaar: Berechnung des Prozentsatzes der kognaten unter allen Wortpaaren als Maß für die Verwandtschaft der beiden Sprachen.

<sup>1</sup> Am einflussreichsten waren hier wohl die Arbeiten von Joseph Greenberg, siehe z.B. Greenberg (1971, 1987, 2000, 2002).

<sup>2</sup> Zwei Wortformen sind *kognat*, wenn sie sich aus derselben Ursprungsform entwickelt haben. So sind etwa das deutsche ‚Wolf‘ und das englische ‚wulf‘ kognat, da beide Formen aus dem urgermanischen ‚\*wulfaz‘ abgeleitet sind. Zwei Wortformen gelten allerdings nicht als kognat, wenn die Verwandtschaft durch Entlehnung vermittelt ist (wie z.B. deutsch ‚Ziegel‘ und italienisch ‚tegola‘; zwar gehen beide Formen auf das lateinische ‚tegula‘ zurück, aber diese Verwandtschaft ist durch eine Entlehnung aus dem Lateinischen ins Althochdeutsche vermittelt).

Jeder dieser Schritte ist mit einer Reihe von Problemen behaftet. Swadesh selbst kompilierte mehrere Versionen einer universalen Konzeptliste, und es wurde nie ein Konsens erzielt, wie eine ideale Liste zusammengesetzt sein soll. Auch gibt es nicht immer genau eine Entsprechung für jedes der fraglichen Konzepte in einer gegebenen Sprache. Für das Swadesh-Konzept *Pfad* etwa gibt es im Deutschen die Entsprechungen ‚Pfad‘ und ‚Weg‘. Weiterhin stellt sich das Problem, ob bei flektierenden Sprachen für Nomina der Nominativ und für Verben der Infinitiv herangezogen werden soll oder lediglich die Wurzel.<sup>3</sup>

Der dritte Schritt ist vermutlich der schwierigste. In vielen Fällen ist die Frage, ob zwei Wörter aus verschiedenen Sprachen kognat sind, grundsätzlich nicht eindeutig zu entscheiden. So listet z.B. die „Indo-European Lexical Cognacy Database“ als Lexikalisierungen (siehe <http://ilex.mpi.nl/>) des Swadesh-Konzepts *Ehefrau* für das Deutsche ‚Frau‘ und für das Färöische (u.a.) ‚húsfrú‘, was unschwer als genaue morphologische Entsprechung des deutschen ‚Hausfrau‘ zu identifizieren ist. Hier handelt es sich also um partielle Kognatheit. Das deutsche Wort ist kognat zu einem Morphem des färöischen Wortes, aber nicht zum gesamten Wortstamm.

Abgesehen von diesem grundsätzlichen Problem ist die Bestimmung der Kognatheit auch unter praktischem Gesichtspunkt schwierig, da ein sicheres Urteil für einen Experten eine Kenntnis der betroffenen Sprachfamilie voraussetzt, also genau die Information, die durch die Lexikostatistik erst gewonnen werden soll. Eine bekannte Illustration dieses Problems ist das Wortpaar (russisch) ‚sto‘ vs. (deutsch) ‚hundert‘. Dank unserer genauen Kenntnisse der Lautverschiebungen, die vom Urindoeuropäischen zum Russischen bzw. zum Deutschen stattgefunden haben, wissen wir, dass sich beide Wörter auf das rekonstruierte ‚\*kmtom‘ zurückführen lassen. Die beiden Wörter sind also kognat. Bei einer weniger gut untersuchten Sprachfamilie würde eine derartige Kognatheitsbeziehung jedoch vermutlich nicht erkannt.

Nicht zuletzt gibt der Prozentsatz der kognaten Wortpaare nur eine recht grobe Schätzung des Grades der Verwandtschaft zwischen zwei Sprachen. Der Grad der Verwandtschaft steht tendenziell in inverser Relation zu der Zeit, die seit der Aufspaltung der gemeinsamen Proto-Sprache verflissen ist.

Dieses Problem lässt sich wiederum anhand der Indo-European Lexical Cognacy Database illustrieren. Der Prozentsatz der als urindoeuropäische Erbwörter ausgewiesenen Einträge pro Sprache variiert zwischen 20% (z.B. für das Paschtunische) und über 50% (für mehrere romanische Spra-

<sup>3</sup> Wenn die Kognatheits-Urteile im dritten Schritt von Experten manuell vorgenommen werden, ist das unproblematisch, aber bei automatischen Verfahren sind diese Entscheidungen durchaus relevant.

chen). Die Ersetzungsrate entlang der verschiedenen Äste der indoeuropäischen Sprachfamilie ist also offensichtlich nicht konstant. Das spiegelt sich auch in den geschätzten Ähnlichkeiten zwischen lebenden Sprachen wieder. In der genannten Datenbank sind 22,5% der Swadesh-Wörter für Spanisch und Hindi kognat, aber nur 14% für Spanisch und Paschtunisch. Wir wissen jedoch mit hoher Sicherheit, dass Hindi und Paschtunisch beide zum indoiranischen Zweig des Indoeuropäischen gehören, Spanisch jedoch zum italischen Zweig, so dass die Zeittiefe seit der letzten gemeinsamen Ursprache für Spanisch/Paschtunisch und Spanisch/Hindi identisch sein muss.

### 3. Elektronisch verfügbare Swadesh-Listen

Mit den Methoden der elektronischen Datenverarbeitung lassen sich Swadesh-Listen wesentlich effizienter und im größeren Maßstab auswerten als zu Swadeshs Lebzeiten.

Die erste elektronisch verfügbare größere Sammlung von Swadesh-Listen war die auf Initiative von Isidore Dyen seit den sechziger Jahren des vorigen Jahrhunderts zusammengestellte „Comparative Indo-European Database“ (erläutert in Dyen/Kruskal/Black 1992). Diese Daten wurden ursprünglich auf Lochkarten kodiert und um 1990 auf elektronische Speichermedien übertragen. Sie umfasst Übersetzungen von 200 Swadesh-Konzepten in 95 indoeuropäische Sprachen und Dialekte sowie Zuordnungen aller Einträge zu Kognatenklassen. Die Wortformen selbst sind in der Orthographie der jeweiligen Sprache angegeben, so dass sich daraus nicht ohne weiteres verlässliche phonetische Informationen gewinnen lassen. Diese Datenbank wird gegenwärtig von der Gruppe „Evolutionary Processes in Language and Culture“ am Max-Planck-Institut für Psycholinguistik Nijmegen unter der Leitung von Michael Dunn als die bereits erwähnte Indo-European Lexical Cognacy Database weitergeführt, umfasst inzwischen 152 Sprachen und Dialekte und enthält für einen Großteil der Einträge auch IPA-Transkriptionen.

Seit einigen Jahren wird von einer Gruppe unter der Leitung von Simon Greenhill von der University of Auckland in Neuseeland die „Austronesian Basic Vocabulary Database“ im Internet zur Verfügung gestellt (siehe Greenhill/Blust/Gray 2008 und die Webseite <http://language.psy.auckland.ac.nz/austronesian/>). Dabei handelt es sich um eine Sammlung von Swadesh-Listen mit jeweils über 200 Einträgen aus (zum gegenwärtigen Zeitpunkt) ungefähr 1000, größtenteils austronesischen, Sprachen. Die Wortformen sind in IPA-Transkriptionen angegeben. Außerdem werden, wie auch in der oben genannten indoeuropäischen Datenbank, von Experten vorgenommene Kognatheitsurteile kodiert.

Einen ähnlichen Umfang hat die Datenbank, die im Rahmen des „Automated Similarity Judgment Program“ (ASJP; siehe Wichmann et al. 2012 bzw. <http://wwwstaff.eva.mpg.de/~wichmann/ASJPHomePage.htm>) unter der Leitung von Søren Wichmann am Max-Planck-Institut für Evolutionäre Anthropologie in Leipzig zusammengestellt wurde. Das ASJP strebt eine repräsentative, möglichst vollständige Erfassung aller lebenden Sprachen und Dialekte an. Gegenwärtig sind ca. 5.600 Sprachvarietäten aus allen Kontinenten und nahezu allen Sprachfamilien erfasst. Das ASJP beschränkt sich dabei auf nur 40 Swadesh-Einträge, die auf der Basis einer kleineren Pilotstudie als besonders stabil identifiziert wurden. Auf die Erhebung von Kognatheitsurteilen wird dabei vollständig verzichtet. Die Wortformen sind in einer einheitlichen phonetischen Transkription kodiert, die im Vergleich zum IPA stark vereinfacht ist. Es werden lediglich 41 verschiedene Segmente unterschieden, die z.T. durch Diakritika modifiziert werden. Um dem Leser einen Eindruck zu vermitteln, sind in Tabelle 1 die ASJP-Listen für das Deutsche und das Englische angegeben. („XXX“ markiert dabei einen fehlenden Eintrag.)

<i>Konzept</i>	Deutsch	Englisch	<i>Konzept</i>	Deutsch	Englisch
<i>ich</i>	iX	Ei	<i>Nase</i>	naz3	nos
<i>du</i>	du	yu	<i>Zahn</i>	ch~an	tu8
<i>wir</i>	vir	wi	<i>Zunge</i>	ch~uN3	t3N
<i>eins</i>	ains	8is	<i>Knie</i>	kni	ni
<i>zwei</i>	cvai	8Et	<i>Hand</i>	hant	hEnd
<i>Mensch</i>	mEnS	pers3n	<i>Brust</i>	brust	brest
<i>Fisch</i>	fiS	fiS	<i>Leber</i>	leb3r	liv3r
<i>Hund</i>	hunt	dag	<i>trinken</i>	triNk3n	drink
<i>Laus</i>	laus	laus	<i>sehen</i>	ze3n	si
<i>Baum</i>	baum	tri	<i>hören</i>	her3n	hir
<i>Blatt</i>	blat	lif	<i>sterben</i>	Sterb3n	dEi
<i>Haut</i>	haut	skin	<i>kommen</i>	kh~om3n	k3m
<i>Blut</i>	blut	bl3d	<i>Sonne</i>	zon3	s3n
<i>Knochen</i>	knoX3n	bon	<i>Stern</i>	StErn	star
<i>Horn</i>	horn	horn	<i>Wasser</i>	vas3r	wat3r
<i>Ohr</i>	XXX	ir	<i>Stein</i>	Stain	ston
<i>Auge</i>	aug3	Ei	<i>Feuer</i>	foia	fEir

Tab. 1: ASJP-Listen für Deutsch und Englisch

Datenmengen in dieser Größenordnung lassen sich selbstredend nicht manuell auswerten. Es ist daher sinnvoll, auf algorithmische Methoden zurückzugreifen, wie sie in den letzten zwei bis drei Jahrzehnten in der Computerlinguistik und der Bioinformatik entwickelt wurden.

## 4. Bioinformatische Methoden für sprachliche Daten

### 4.1 Phylogenetische Inferenz

Sowohl die empirische Basis des lexikostatistischen Vorgehens wie auch die gewonnenen Ergebnisse sind also mit großer Unsicherheit behaftet. Diese Situation ist aber für datenorientierte Arbeit generell nicht ungewöhnlich und für sich genommen kein Grund, die Methode zu verwerfen. Die größte Schwäche der klassischen Lexikostatistik ist m.E. ironischerweise die Tatsache, dass sie zwar quantitativ arbeitet, jedoch auf *statistische* Methoden im engeren Sinne verzichtet. Für statistisches Arbeiten ist es gerade typisch, auf der Basis unsicherer Daten unsichere Inferenzen zu ziehen, jedoch den Grad der Unsicherheit der Ergebnisse zu quantifizieren.

Angenommen, wir haben für eine Gruppe von Sprachen mit lexikostatistischen Methoden die paarweisen Ähnlichkeiten gewonnen. Daraus lassen sich zwar nicht mit Sicherheit die zeitlichen Abstände zwischen zwei beliebigen Sprachen berechnen, aber es ist zu erwarten, dass diese Ähnlichkeiten negativ mit den tatsächlichen Abständen korreliert sind. Eine bestimmte Hypothese über die Verwandtschaftsverhältnisse, also ein hypothetischer Sprachstammbaum, erklärt die berechneten Ähnlichkeiten umso besser, je stärker diese mit den angenommenen zeitlichen Abständen korrelieren. Daher kann man umgekehrt von den berechneten Ähnlichkeiten ausgehen und eine Hypothese suchen, die diese Werte am besten erklärt.

Strukturell ähnliche Probleme wurden in den letzten zwanzig bis dreißig Jahren in der Bioinformatik gründlich untersucht. Die Bioinformatik befasst sich u.a. mit der Frage, wie die evolutionäre Geschichte von Organismen mit algorithmischen und statistischen Methoden rekonstruiert werden kann.

Auch in der Biologie ist die Ähnlichkeit zweier Organismen oder Populationen – die entweder über geteilte phänotypische oder genetische Merkmale bestimmt werden kann – ein Hinweis auf die zeitliche Distanz zum letzten gemeinsamen Vorfahren im Verlauf der Evolution, und auch in der Biologie ist diese Abschätzung häufig nur näherungsweise möglich. Ein besonders gut untersuchtes Problem ist die Frage, wie aus einer Ähnlichkeitsmatrix ein Stammbaum berechnet werden kann, der diese Ähnlichkeitsmatrix optimal erklärt. Es ist zwar nicht möglich, mit realistischem Rechenaufwand den besten Stammbaum zu ermitteln,<sup>4</sup> aber es existieren sehr gute Näherungsverfahren. Im Folgenden werde ich mich dazu der ‚Fastme‘-Methode (Desper/Gascuel 2002) bedienen.<sup>5</sup>

<sup>4</sup> Dieses Problem ist NP-vollständig, also für größere Datenmengen praktisch nicht lösbar.

<sup>5</sup> Dabei wird zunächst mit Hilfe des Neighbor-Joining-Algorithmus (Saitou/Nei 1987) oder eines ähnlichen distanzbasierten Verfahrens ein phylogenetischer Baum berechnet und die

## 4.2 Sequenzalinierung

Ein wesentlicher Schritt der lexikostatistischen Methode ist die Erhebung von Kognatheitsurteilen. Wie oben ausgeführt, ist dieser Schritt nicht unproblematisch. Er muss manuell von Experten vorgenommen werden, die ihre Urteile wiederum auf Theorien über die historische Verwandtschaft der zu vergleichenden Sprachen stützen. Daher ist es unvermeidlich, dass es bei weniger gut untersuchten Sprachfamilien einen größeren Anteil an ‚false negatives‘ gibt als z.B. beim Indoeuropäischen. Auch liegen relativ unkontroverse Kognatheitsurteile im ausreichendem Umfang bislang lediglich für das Indoeuropäische und das Austronesische vor. Daher werde ich diesen Schritt durch ein automatisiertes Verfahren ersetzen, das die Ähnlichkeit<sup>6</sup> zweier Wortformen aus der ASJP-Datenbank quantifiziert. Dieses Vorgehen ist mit Joseph Greenbergs ‚lexical mass comparison‘ verwandt. Im Unterschied zu Greenberg ist das von mir verwendete Ähnlichkeitsmaß jedoch klar definiert, so dass meine Methode vollständig reproduzierbar ist.

Die vermutlich einfachste Methode, die Ähnlichkeit zweier Symbolketten zu quantifizieren, basiert auf der sogenannten ‚Levenshtein-Distanz‘ (bzw. Editier-Distanz). Die Levenshtein-Distanz zweier Ketten  $k_1$  und  $k_2$  ist die minimale Zahl von Editieroperationen (also Einfügen, Tilgen oder Ersetzen eines einzelnen Symbols), die  $k_1$  in  $k_2$  überführt. Das sei anhand der ASJP-Einträge (dt.) *horn* und (lat.) *kornu* (für das Konzept *Horn*) illustriert. Es sind zwei Editierschritte nötig: Ersetzung von h durch k und Einfügung des u.

Die Levenshtein-Distanz lässt sich auch als die Zahl der Nicht-Übereinstimmungen in der optimalen Alinierung der betreffenden Symbolketten auffassen. Das ist in Abbildung 1 illustriert.

```

h o r n
: | | | :
k o r n u

```

Abb. 1: Levenshtein-Alinierung

Die ‚normalisierte Levenshtein-Distanz‘ ergibt sich, wenn man diese Distanz durch die Länge der längeren Kette teilt. Im Beispiel ergibt sich dabei ein Wert von 0,4.

---

ser dann in einem zweiten Schritt durch ‚nearest neighbor interchange‘ so lange lokal optimiert, bis keine Verbesserung mehr möglich ist.

<sup>6</sup> Im Folgenden setze ich stillschweigend voraus, dass sich ein Ähnlichkeitsmaß leicht in ein Distanzmaß umrechnen lässt und umgekehrt.

Die paarweise Distanz zweier Wortlisten lässt sich jetzt leicht als die durchschnittliche Distanz zwischen ihren korrespondierenden Einträgen definieren.

Wenn man mit dieser Methode die paarweisen Distanzen der westgermanischen Sprachen und Dialekte aus der ASJP-Datenbank berechnet und daraus mit Hilfe des *fastme*-Verfahrens einen phylogenetischen Baum gewinnt, erhält man das in Abbildung 2 gezeigte Ergebnis. Dieser Stammbaum stimmt mit einer Expertenklassifikation zwar nicht in jedem Detail überein, liefert aber eine recht gute Annäherung.

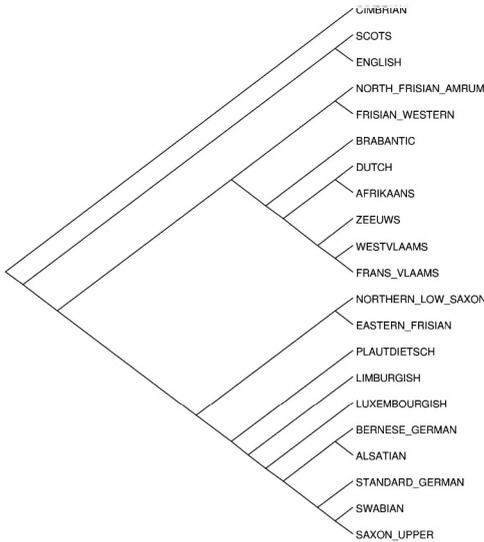


Abb. 2: Stammbaum der westgermanischen Sprachen: einfache Levenshtein-Alinierung

Wenn man dieses Verfahren jedoch auf die gesamte ASJP-Datenbank<sup>7</sup> anwendet, zeigt sich allerdings schnell ein schwerer Defekt. Wenn zwei Sprachen kleine Lautinventare haben, die sich überlappen, ergeben sich eine größere Anzahl von zufälligen Übereinstimmungen bei der Alinierung als beim Vergleich von Sprachen mit großen oder nicht so stark überlappenden Lautinventaren. Das lässt sich anhand der Grafik in Abbildung 3 erläutern. Diese Grafik wurde mit Hilfe der Software CLANS (Frickey/Lupas 2004) erstellt. Jeder Kreis repräsentiert eine Sprache. Sprachen mit geringer Distanz entsprechen nahe beieinanderliegenden Punkten und vice versa. Die Punkte sind entsprechend der Zuordnung der jeweiligen Sprache zu Sprachfamilien nach dem „World Atlas of Language Structures“ (WALS; siehe

<sup>7</sup> Genauer gesagt: auf die lebenden oder kürzlich ausgestorbenen Sprachen und Dialekte in der Datenbank unter Ausschluss der Kreolsprachen.

Haspelmath et al. 2008) in Graustufen eingefärbt. Es ist leicht zu sehen, dass es im Zentrum der Grafik eine große Zahl von Sprachen aus verschiedenen Sprachfamilien gibt, die eine geringe Distanz zueinander haben. Eine genauere Inspektion der Daten ergab, dass es sich dabei in der Tat um Sprachen mit kleinem Lautinventar handelt, diese Ähnlichkeiten also nicht auf genetischer Verwandtschaft beruhen.

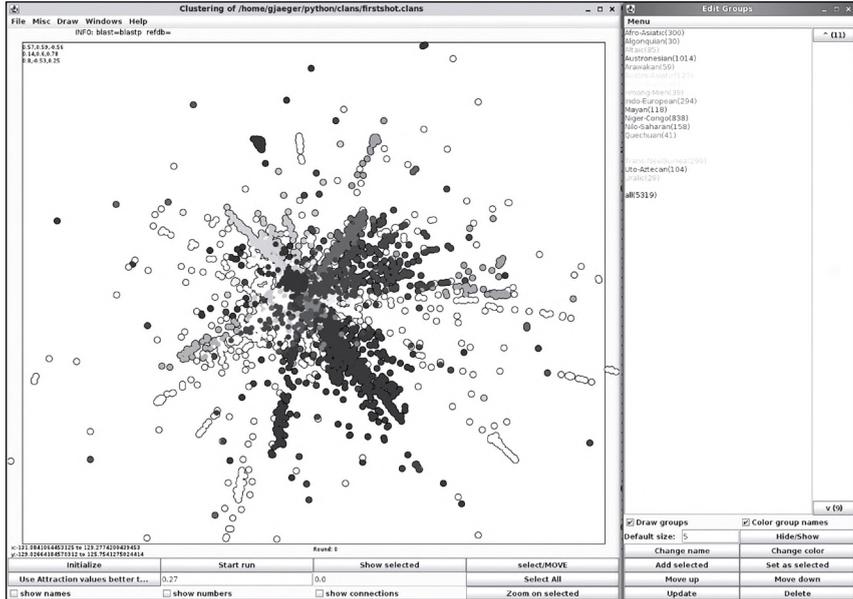


Abb. 3: Visualisierung der Ähnlichkeiten in ASJP: einfache Levenshtein-Alinierung

Um diesen Effekt zu neutralisieren, ist es nötig, die Distanz zwischen zwei Wortformen für die phonetischen Charakteristika der verglichenen Sprachen zu kalibrieren.

Das hierbei angewandte Verfahren sei anhand des Vergleichs von Englisch und Schwedisch illustriert.<sup>8</sup> Im ersten Schritt werden die normalisierten Levenshtein-Distanzen für alle 1.600 Wortpaare aus den beiden Sprachen berechnet. Ein Teil der so gewonnenen 40×40-Matrix ist in Tabelle 2 dargestellt. Die Einträge entlang der Diagonale geben die Distanzen zwischen synonymen Wortformen wieder. Die restlichen Einträge stellen eine Stichprobe der Verteilung von Distanzen dar, die zwischen zufällig gewählten, nicht verwandten englisch-schwedischen Wortpaaren bestehen. Je stärker zwei Sprachen miteinander verwandt sind, umso mehr sollten sich die Verteilung der Werte auf der Diagonale und die Verteilung der restlichen

<sup>8</sup> Eine detailliertere Darstellung des im folgenden skizzierten Verfahrens findet sich in Jäger (2013).

Werte unterscheiden. Dabei ist zu erwarten, dass bei verwandten Sprachen die Diagonal-Einträge deutlich kleiner sind als die anderen Einträge. Für das Sprachpaar Englisch-Schwedisch ist das in der Tat der Fall. Die beiden Verteilungen sind in der Grafik links in Abbildung 4 dargestellt.

	Ei	yu	wi	w3n	tu	fiS	...
yog	<b>1</b>	$\frac{2}{3}$	1	1	1	1	
du	1	$\frac{1}{2}$	1	1	$\frac{1}{2}$	1	
vi	$\frac{1}{2}$	1	$\frac{1}{2}$	1	1	$\frac{2}{3}$	
et	1	1	1	<b>1</b>	1	1	
tvo	1	1	1	1	$\frac{2}{3}$	1	
fisk	$\frac{3}{4}$	1	$\frac{3}{4}$	1	1	$\frac{1}{2}$	
:							

Tab. 2: Normalisierte Levenshtein-Distanzen: Englisch/Schwedisch 2/3

Beim Vergleich der nicht verwandten Sprachen Englisch und Türkisch ergibt sich im Kontrast dazu, dass die Diagonalwerte im Schnitt sogar etwas größer sind als die restlichen Werte, synonyme Wortpaare sich also sogar etwas stärker voneinander unterscheiden als Zufallswortpaare (siehe rechte Grafik in Abbildung 4).

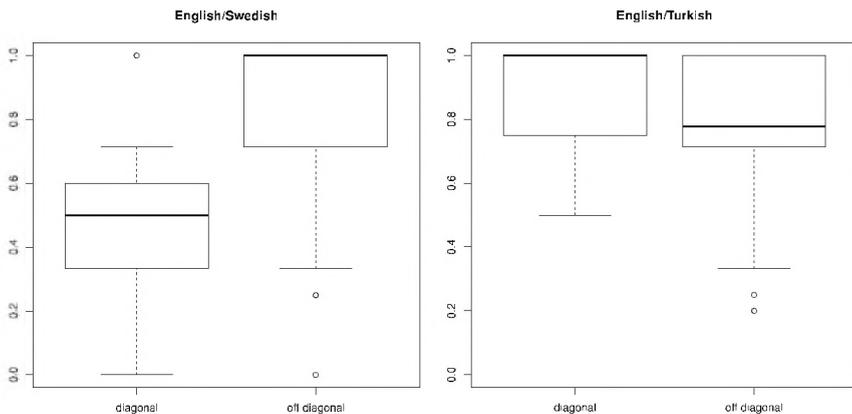


Abb. 4: Verteilung von Levenshtein-Distanzen

Mit Hilfe eines nicht-parametrischen statistischen Tests lässt sich die Wahrscheinlichkeit abschätzen, dass die Diagonalwerte in einer solchen Matrix derselben Verteilung entstammen wie die restlichen Werte. Diese Wahrscheinlichkeit (in statistischer Terminologie:  $p$ -Wert) gibt ein inverses Maß für den Grad der Verwandtschaft der verglichenen Sprachen. Für Englisch-Schwe-

disch beträgt dieser Wert ungefähr  $10^{-70}$ , für Englisch-Türkisch 0,67. Es ist also praktisch ausgeschlossen, dass die Diagonalverteilung für Englisch-Schwedisch zufällig so stark von der sonstigen Verteilung abweicht, während das Muster bei Englisch-Türkisch dem entspricht, was man bei einer Zufallsverteilung erwartet.

Aus diesen  $p$ -Werten werden durch nicht-lineare Transformationen Ähnlichkeitsmaße gewonnen, die die Basis für phylogenetische Inferenz liefern.

In Tabelle 3 sind die so berechneten Ähnlichkeiten des Standard-Deutschen zu einer Reihe ausgewählter Sprachen und Dialekte aufgeführt.

Schwäbisch	26,13
Zimbrisch	20,28
Niederländisch	23,75
Englisch	17,45
Urindoeuropäisch	10,26
Latein	9,23
Spanisch	8,95
Hindi	8,70
Russisch	8,36
Türkisch	6,33
Ungarisch	6,84

Tab. 3: Ähnlichkeiten zum Standard-Deutschen: kalibrierte Levenshtein-Alinierung

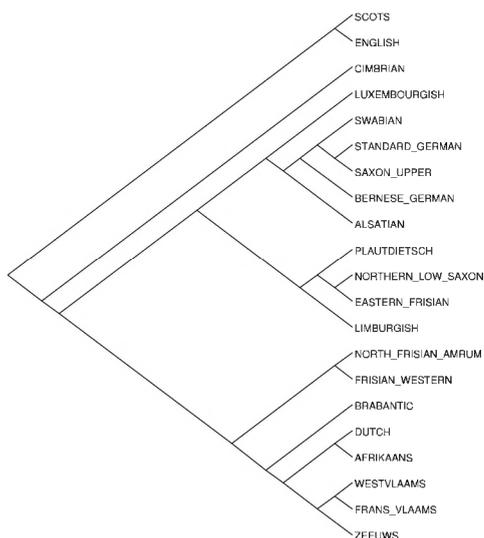


Abb. 5: Stammbaum der westgermanischen Sprachen: kalibrierte Alinierung

Um das Illustrationsbeispiel der westgermanischen Sprachgruppe weiterzuführen, ist in Abbildung 5 der auf der Basis der kalibrierten Levenshtein-Distanzen berechnete Stammbaum dargestellt. Dieser Stammbaum unterscheidet sich nicht wesentlich von dem in Abbildung 3 gezeigten.

Ein offensichtlicher Mangel dieser Stammbäume ist die Tatsache, dass das Zimbrische – ein im Trentino gesprochener bairischer Dialekt – ganz an der Peripherie der westgermanischen Sprachen angesiedelt ist, obwohl es sich dabei um einen oberdeutschen Dialekt handelt. Es ist erhellend, die zimbri-sche ASJP-Liste genauer zu betrachten. Sie ist in Tabelle 4 wiedergegeben.

<i>Konzept</i>	Deutsch	Zimbrisch	<i>Konzept</i>	Deutsch	Zimbrisch
<i>ich</i>	iX	ix	<i>Nase</i>	naz3	naza
<i>du</i>	du	du	<i>Zahn</i>	ch~an	XXX
<i>wir</i>	vir	bar	<i>Zunge</i>	ch~uN3	suNa
<i>eins</i>	ains	XXX	<i>Knie</i>	kni	XXX
<i>zwei</i>	cvai	sben	<i>Hand</i>	hant	hant
<i>Mensch</i>	mEnS	menEs	<i>Brust</i>	brust	prust
<i>Fisch</i>	fiS	XXX	<i>Leber</i>	leb3r	lEbara
<i>Hund</i>	hunt	hunt	<i>trinken</i>	triNk3n	trinkh~
<i>Laus</i>	laus	laus	<i>sehen</i>	ze3n	zeg
<i>Baum</i>	baum	pom	<i>hören</i>	her3n	hor
<i>Blatt</i>	blat	placa	<i>sterben</i>	Sterb3n	sterb
<i>Haut</i>	haut	XXX	<i>kommen</i>	kh~om3n	kh~Em
<i>Blut</i>	blut	plut	<i>Sonne</i>	zon3	zuna
<i>Knochen</i>	knoX3n	poan	<i>Stern</i>	StErn	stErna
<i>Horn</i>	horn	horn	<i>Wasser</i>	vas3r	basar
<i>Ohr</i>	XXX	oar	<i>Stein</i>	Stain	stoan
<i>Auge</i>	aug3	ogh~E	<i>Feuer</i>	foia	boar

Tab. 4: ASJP-Listen für Standard-Deutsch und Zimbrisch

Es fällt auf, dass im Zimbrischen eine Reihe von regulären Lautverschiebungen stattgefunden haben, von denen die meisten hochdeutschen Dialekte nicht betroffen sind. So finden wir nicht nur die konsequente Anwendung der zweiten Lautverschiebung auf b, das zu p wird (baum-pom, blat-placa, blut-plut). Auffällig ist vor allem die ungewöhnliche Verschiebung von v (entspricht dem ‚w‘ in der deutschen Orthographie) zu b: vir-bar, cvai-sben, vas3r-basar.

Diese Lautkorrespondenzen sind für einen geschulten Linguisten natürlich unschwer zu erkennen, da sie artikulatorisch völlig natürlich sind. Die Levenshtein-Alinierung unterscheidet jedoch nur zwischen identischen und nicht-identischen Segmenten. Daher erscheint das Zimbrische weiter von den anderen hochdeutschen Dialekten entfernt, als es tatsächlich ist.

Anhand des Paares *blat-placa* (gesprochen ‚Platza‘; das ASJP-Symbol *c* steht für die dentale Affrikate) (Standard-Deutsch bzw. Zimbrisch für *Blatt*) lässt sich dieses Problem näher beleuchten. Hier liegt eine fast vollständige 1-1-Korrespondenz der einzelnen Segmente vor. Die Korrespondenzen *v-b* und *t-c* sind völlig regulär und sind Evidenz für, nicht gegen die Annahme, dass die beiden Wörter kognat sind. Die normalisierte Levenshtein-Distanz beträgt jedoch 0,6, ein relativ hoher Wert. Für das nicht kognate Wortpaar (dt.) *hunt* ‚Hund‘ – (zimbr.) *zuna* ‚Sonne‘ (das ASJP-Symbol *z* symbolisiert ein stimmhaftes *S*) ergibt sich z.B. eine geringere Distanz von 0,5.

Ein vergleichbares Problem stellt sich in der Bioinformatik, wenn Proteinsequenzen aliniert werden. Idealerweise sollten solche Aminosäuren einander zugeordnet werden, die auf einen gemeinsamen evolutionären Vorfahren zurückgehen und ggf. durch Mutationen verändert wurden. Allerdings sind nicht alle Ersetzungen von Aminosäuren durch Mutationen gleich wahrscheinlich. Die beste Alinierung ist daher diejenige, die die Wahrscheinlichkeit maximiert, dass einander zugeordnete Positionen evolutionär verwandt sind.<sup>9</sup>

Dazu werden für jedes Paar von Aminosäuren die *odds* bestimmt, dass sie evolutionär verwandt sind. Die *odds* sind der Quotient aus der Wahrscheinlichkeit, dass die betreffenden Säuren durch Mutationen aus demselben Vorfahren hervorgegangen sind, und der Wahrscheinlichkeit, dass sie zufällig in nicht verwandten Sequenzen einander zugeordnet werden. Die optimale Alinierung zweier Sequenzen ist die, die das Produkt dieser punktweisen *odds* maximiert.

Üblicherweise arbeitet man mit den *log-odds*, also den Logarithmen der *odds*. Die optimale Alinierung maximiert dann die Summe der einzelnen *log-odds*. Diese Summe ist ein Maß dafür, wie plausibel die Annahme ist, die beiden Sequenzen seien verwandt. Positive Werte bedeuten dabei, dass die Evidenz für eine Verwandtschaft überwiegt, und negative Werte deuten entsprechend auf ein Überwiegen der Evidenz gegen eine Verwandtschaft hin.

Die optimale Alinierung lässt sich mit Hilfe des ‚Needleman-Wunsch-Algorithmus‘ (Needleman/Wunsch 1970) effizient berechnen.

Analog ist es auch für verschiedene Lautpaare unterschiedlich wahrscheinlich, dass sie durch reguläre Lautverschiebungen auf eine gemeinsame Urform zurückgehen. Wie in Jäger (2013) dargestellt, lassen sich die entsprechenden *odds* durch Heuristiken anhand der ASJP-Daten abschätzen.

Für die Zuordnung *b-p* betragen die geschätzten *log-odds* 0,46, eine solche Korrespondenz ist also als (schwache) Evidenz für eine etymologische Verwandtschaft der entsprechenden Wörter zu werden. Der Wert für *t-c* ist 0,08, also ebenfalls leicht positiv.

<sup>9</sup> Für eine ausführliche Darstellung der bioinformatischen Methoden der Sequenzalinierung siehe z.B. Durbin et al. (1989).

Bei der Alinierung von *hunt* und *zuna* werden *h* und *z* einander zugeordnet. Die *log-odds* dafür sind mit  $-0,91$  deutlich negativ, wie auch die Alinierung *t-a* mit  $-8,14$ . Die aggregierten *log-odds* für das Wortpaar *blat-placa* betragen  $4,05$ , während *hunt-zuna* mit  $-3,76$  bewertet wird. Dieses Beispiel illustriert, dass die Abschätzung von Wortähnlichkeiten via *log-odds* wesentlich besser geeignet ist als die normalisierte Levenshtein-Distanz, um kognate Wortpaare von nicht-kognaten zu unterscheiden.

	ungewichtet	gewichtet
Schwäbisch	26,13	35,44
Zimbrisch	20,28	31,86
Niederländisch	23,75	29,76
Englisch	17,45	22,14
Urindoeuropäisch	10,26	15,86
Latein	9,23	12,54
Spanisch	8,95	9,48
Hindi	8,70	12,35
Russisch	8,36	11,89
Türkisch	6,33	5,76
Ungarisch	6,84	7,57

Tab. 5: Ähnlichkeiten zum Standard-Deutschen: Levenshtein-Alinierung vs. gewichtete Alinierung

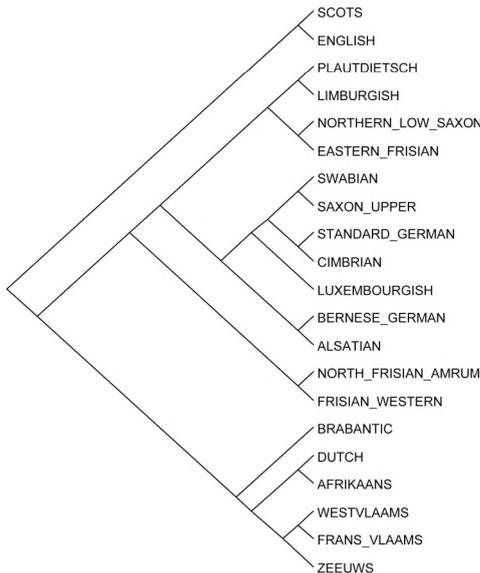


Abb. 6: Stammbaum der westgermanischen Sprachen: kalibrierte gewichtete Alinierung

Die Berechnung der kalibrierten Ähnlichkeit zweier Sprachen auf der Basis der aggregierten *log-odds* kann dann analog zum entsprechenden Vorgehen mit Levenshtein-Distanzen erfolgen.

Es ist instruktiv, die auf der Basis der *log-odds* abgeschätzten Ähnlichkeitswerte für die in Tabelle 3 aufgeführten Beispielsprachen/-dialekte mit den dort gegebenen Werten zu vergleichen (vgl. Tabelle 5). Die enge Verwandtschaft des Zimbrischen zum Standard-Deutschen, im Vergleich etwa zum Niederländischen, wird mit der neuen Methode korrekt erkannt. Dementsprechend ist der auf der Basis der gewichteten Alinierung gewonnene Stammbaum (Abbildung 6) genauer als die bisher betrachteten Versionen. Insbesondere wird das Zimbrische hier korrekt als Teil der hochdeutschen Dialektgruppe identifiziert.

## 5. Anwendungen auf größere Datenmengen

Das im vorherigen Abschnitt dargestellte Verfahren wurde auf eine ausgewählte Teilmenge der (lebenden oder erst kürzlich ausgestorbenen) Sprachen und Dialekte in der ASJP-Datenbank angewandt. Ausgewählt wurden alle europäischen und asiatischen Sprachen (mit den unten genannten Ausnahmen) einschließlich der in Afrika gesprochenen afro-asiatischen Sprachen. Außerdem wurden die amerikanischen eskimo-aleutischen und Na-Dené-Sprachen sowie die austronesischen Sprachen in die Auswahl aufgenommen. Die Auswahl begründet sich damit, dass verschiedentlich in der Literatur vorgeschlagen wurde, es gebe tiefe genetische Beziehungen der afro-asiatischen, eskimo-aleutischen oder Na-Dené-Sprachen zu europäischen bzw. asiatischen Sprachen. Nach der populären, aber kontroversen, *nostratischen* Hypothese (siehe z.B. Bomhard/Kerns 1994) bilden Afro-Asiatisch, Indoeuropäisch, Uralisch, Altaisch, Kartwelisch, Jukagirisch, Eskimo-Aleutisch, Tschuktscho-Kamtschadalisch und möglicherweise Dravidisch eine Makro-Familie. Weiterhin wurde (etwa in Nikolaev 1991) eine tiefe Verwandtschaft zwischen den nordkaukasischen Sprachen und den Na-Dené-Sprachen postuliert. Da eine Verwandtschaft von Na-Dené mit den sino-tibetischen Sprachen schon verschiedentlich angenommen wurde (u.a. in unveröffentlichten Arbeiten von Edward Sapir; siehe Campbell/Poser 2008), wird teilweise auch angenommen, dass die Na-Dené-Sprachen gemeinsam mit Sino-Tibetisch, möglicherweise Burushaski, den nordkaukasischen Sprachen und eventuell auch Baskisch eine Makro-Familie bilden. Nicht zuletzt gibt es eine Reihe von Vorschlägen, die das Austronesische mit südostasiatischen Sprachen in Beziehung setzen, so z.B. der Vorschlag von Benedict (1975), dass Austronesisch und Tai-Kadai eine Makro-Familie namens Austro-Tai bilden.

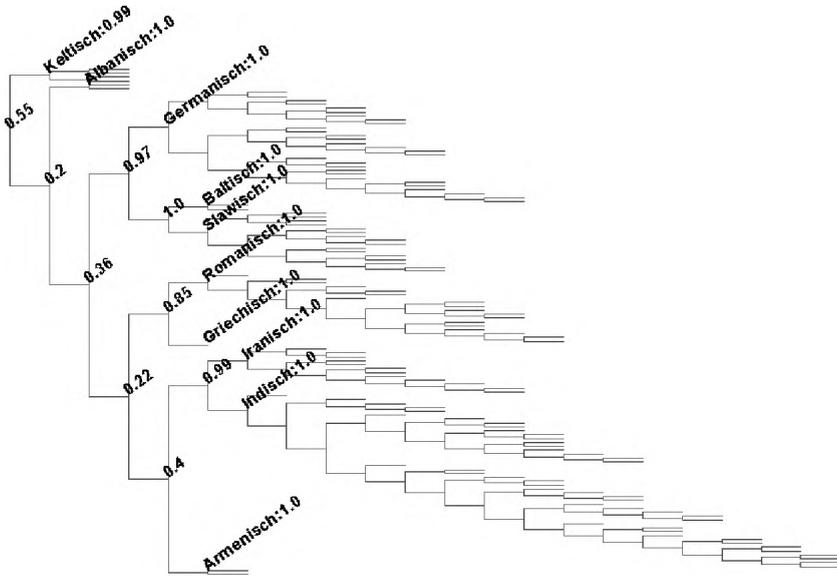


Abb. 7: Automatisch erstellter Stammbaum der indoeuropäischen Sprachen

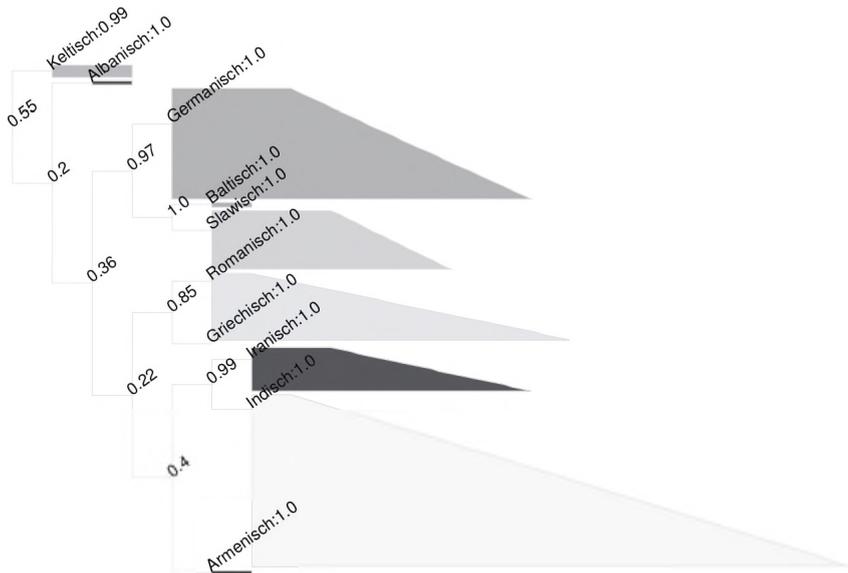


Abb. 8: Automatisch erstellter Stammbaum der indoeuropäischen Sprachen: schematische Darstellung

Die isolierten Sprachen Baskisch, Ainu, Koreanisch, Shompen, Nahali, und Kusunda wurden ausgeklammert, da dafür nur jeweils eine Wortliste vorliegt und diese geringe Datenmenge keine statistisch aussagekräftigen Schlüsse zulassen. Gleichfalls ausgeschlossen wurden Japanisch und die dravidische Sprache Brahui, da die entsprechenden ASJP-Wortlisten ungewöhnlich viele klar identifizierbare Lehnwörter enthalten (beim Japanischen aus dem Chinesischen, bei Brahui aus indo-arischen Sprachen), die das gewonnene Bild verfälschen.

Insgesamt enthält die verwendete Auswahl 1.323 Wortlisten. Daraus wurde mit der im letzten Abschnitt dargestellten Methode automatisch ein Stammbaum erstellt.

Der die indoeuropäischen Sprachen umfassende Teilbaum ist in Abbildung 7, und in einer schematischen Darstellung in Abbildung 8 dargestellt. Bemerkenswerterweise werden die etablierten Untergruppen des Indoeuropäischen ausnahmslos korrekt erkannt – keine einzige Sprache wird falsch klassifiziert. Auch die weitgehend akzeptierten größeren Gruppierungen *Indo-Iranisch* und *Balto-Slawisch* werden erkannt.

In welcher Beziehung diese Untergruppen zueinander stehen, ist seit dem 19. Jahrhundert in der Indoeuropäistik kontrovers. Für einige in dem automatisch generierten Stammbaum vorgeschlagene Strukturen, wie z.B. die enge Verbindung des Germanischen mit dem Balto-Slawischen gibt es entsprechende Vorschläge in der Literatur (z.B. Schleicher 1861). Um die Verlässlichkeit derartiger Hypothesen abzuschätzen, wurde eine statistische Analyse vorgenommen. Zu der automatisch gewonnenen Distanzmatrix wurde 1.000 mal zufällig verteilte kleine Rauschwerte addiert und aus den verrauschten 1.000 Matrizen jeweils ein phylogenetischer Baum berechnet. Für jede Verzweigung im Referenzbaum wurde bestimmt, wie häufig die entsprechende Gruppierung in den 1.000 Varianten vorkommt. Die in den Abbildungen angegebenen Zahlen geben die relativen Häufigkeiten. Diese Werte sind also als Maß für die Konfidenz der jeweiligen Gruppierung zu werten.

Für alle etablierten Untergruppen besteht eine Konfidenz von nahezu 100%. Auch für das Balto-Slawische ist die Konfidenz 100% und für das Indo-Iranische 99%. Die anderen höheren Gruppierungen haben alle eine geringere Konfidenz, mit der Ausnahme der Zusammenfassung von Balto-Slawisch und Germanisch in eine Gruppe. Dieser Effekt könnte allerdings auch auf jahrhundertelangen Sprachkontakt zurückzuführen sein.

Der Stammbaum für die gesamte untersuchte Auswahl an Sprachen und Dialekten ist in Abbildung 9 schematisch dargestellt. Auch hier stimmt die automatisch erzielte Klassifikation gut mit der üblichen Expertenklassifikation überein. Es gibt insgesamt nur drei Abweichungen von der WALS-Klassifikation in Sprachfamilien:



terweise bilden auch die Na-Dené-Sprachen und die kaukasischen Sprachen (unter Einschluss des Ket) eine Einheit, mit Konfidenz von 11%. Es gibt allerdings keine Evidenz dafür, dass diese Einheit einer größeren dené-sino-kaukasischen Makrofamilie wäre. Die dené-kaukasische Einheit ist die Gruppierung, die als erste von der Wurzel des Baumes (symbolisiert durch den weißen Kreis in der Mitte der Grafik) abzweigt.

Die hypothetische nostratische Makrofamilie wird – unter Ausschluss der drawidischen Sprachen – ebenfalls als Einheit dargestellt, allerdings mit sehr geringer Konfidenz von 4%.

## 6. Zusammenfassung

Das primäre Ziel dieser Arbeit war es, zu demonstrieren, dass Lexikostatistik linguistisch belastbare Ergebnisse liefert, wenn sie mit modernen – das heißt: computergestützten und statistischen – Mitteln betrieben wird. In der hier vorgestellten Fallstudie werden die traditionellen Einheiten der Sprachklassifikation weitgehend korrekt repliziert. Tendenziell ist es so, dass solche Einheiten, die durch die komparative Methode sicher demonstriert werden können, auch mit einer hohen Konfidenz erkannt werden. Darüber hinaus finden sich einige der kontroverseren Vorschläge für tiefe genetische Beziehungen zwischen Sprachen, wie Nostratisch oder Dené-Kaukasisch, in der automatischen Klassifikation wieder, allerdings mit wesentlich geringerer Konfidenz. Dieser Befund deutet darauf hin, dass die Lexikostatistik letztendlich aus ähnlichen Daten, wie sie in der traditionell-komparativen Klassifikation verwendet werden, ähnliche Schlüsse zieht, auch wenn die Art der Inferenz eine andere ist.

## Literatur

- Benedict, Paul (1975): *Austro-Thai language and culture; with a glossary of roots*. New Haven.
- Bomhard, Allan R./Kerns, John C. (1994): *The Nostratic macrofamily: a study in distant linguistic relationship*. (= *Trends in Linguistics: Studies and Monographs* 74). Berlin/New York.
- Campbell, Lyle/Poser, William J. (2008): *Language classification: history and method*. Cambridge.
- Desper, Richard/Gascuel, Olivier (2002): *Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle*. In: *Journal of Computational Biology* 9, S. 687–705.
- Durbin, Richard et al. (1998): *Biological Sequence Analysis: probabilistic models of proteins and nucleic acids*. Cambridge, UK.

- Dyen, Idiosodre/Kruskal, Joseph B./Black, Paul (1992): An Indoeuropean classification: A lexicostatistical experiment. In: *Transactions of the American Philosophical Society* 82, 5, S. 1–132.
- Frickey, Tancred/Lupas, Andrei N. (2004): CLANS: a Java application for visualizing protein families based on pairwise similarity. In: *Bioinformatics* 20, S. 3702–3704.
- Greenberg, Joseph H. (1971): The Indo-Pacific hypothesis. In: Sebeok, Thomas A. (Hg): *Current trends in linguistics*. Bd. 8: *Linguistics in Oceania*. Den Haag, S. 809–871.
- Greenberg, Joseph H. (1987): *Language in the Americas*. Stanford, CA.
- Greenberg, Joseph H. (2000): *Indo-European and its closest relatives: the Eurasiatic language family*. Bd. 1: *Grammar*. Stanford, CA.
- Greenberg, Joseph H. (2002): *Indo-European and its closest relatives: the Eurasiatic language family*. Bd. 2: *Lexicon*. Stanford, CA.
- Greenhill, Simon J./Blust, Robert/Gray, Russell D. (2008): The Austronesian Basic Vocabulary Database: from bioinformatics to lexomics. In: *Evolutionary Bioinformatics* 4, S. 271–283.
- Haspelmath, Martin et al. (2008): *The World Atlas of Language Structures Online*. Max Planck Digital Library. München. Internet: <http://wals.info/> (Stand: 6.8.2013).
- Jäger, Gerhard (2013): *Phylogenetic inference from word lists using weighted alignment with empirically determined weights*. Ms. Universität Tübingen/Swedish Collegium of Advanced Study Uppsala.
- Needleman, Saul B./Wunsch, Christian D. (1970): A general method applicable to the search for similarities in the amino acid sequence of two proteins. In: *Journal of Molecular Biology* 48, S. 443–453.
- Nikolaev, Sergei (1991): *Sino-caucasian languages in America*. Ms.
- Nikolaev, Sergei/Starostin, Sergei (1994): *The North Caucasian Etymological Dictionary*. Moskau.
- Saitou, Naruya/Nei, Masatoshi (1987): The neighbor-joining method: a new method for reconstructing phylogenetic trees. In: *Molecular Biology and Evolution* 4, S. 406–425.
- Schleicher, August (1861): *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*. Bd. 1. Weimar.
- Swadesh, Morris (1971): *The origin and diversification of language*. Chicago.
- Vajda, Edward J. (2010): A Siberian link with Na-Dene languages. The Dene-Yeniseian connection. In: Kari, James/Potter, Ben A. (Hg): *The Dene-Yeniseian connection*. (= *Anthropological Papers of the University of Alaska* 5). Fairbanks, S. 33–99.
- Wichmann, Søren et al. (2012): *The ASJP Database (Version 15)*. Internet: <http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm> (Stand: 6.8.2013).