Wie die Bioinformatik hilft, Sprachgeschichte zu rekonstruieren

Gerhard Jäger

Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmstraße 19
72074 Tübingen , Deutschland

Swedish Collegium for Advanced Study
Thunbergsvägen 2
752 38 Uppsala , Schweden

> *"The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. [...] We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation. The manner in which certain letters or sounds change when others change is very like correlated growth. [...] The frequent presence of rudiments, both in languages and in species, is still more remarkable. [...] Languages, like organic beings, can be classed in groups under groups; and they can be classed either naturally according to descent, or artificially by other characters. Dominant languages and dialects spread widely, and lead to the gradual extinction of other tongues." (aus Charles Darwin, The Descent of Man, 1871)*

1. Einführung: Linguistik und Biologie

In dem Eingangszitat weist Darwin auf bemerkenswerte Parallelen zwischen biologischer Evolution und dem Wandel natürlicher Sprachen hin. Im neunzehnten Jahrhundert gab es deshalb vielfältige wechselseitige Inspirationen zwischen Evolutionsbiologie und historischer Linguistik. Während des zwanzigsten Jahrhunderts konzentrierte sich die Sprachforschung auf die synchrone Struktur einzelner Sprachen, so dass der evolutionäre Aspekt in den Hintergrund trat. In den letzten zehn bis fünfzehn Jahren ist das Interesse an Sprachevolution jedoch neu erwacht. Die atemberaubenden Fortschritte in der Evolutionsbiologie und in der Bioinformatik der letzten Jahrzehnte liefern den Linguisten

Forschungswerkzeuge, die einen völlig neuen Zugang zu Fragen gestatten, deren Beantwortung noch vor wenigen Jahrzehnten als aussichtslos erschien.

Der Vergleich biologischer Spezies legt eine taxonomische Ordnung nahe. So gruppieren sich die Tiere in Wirbeltiere und Wirbellose, die Wirbeltiere umfassen (u.a.) Fische, Vögel, Reptilien und Säugetiere, die Säugetiere umfassen u.a. Raubtiere, Primaten usw. Schon Aristoteles entwickelte eine hierarchische Klassifikation des Tierreichs (siehe Atkinson & Gray 2005:514). In der Arbeit Carl von Linnés (1707-1778) fand die hierarchische Klassifikation des Tier- und Pflanzenreichs einen vorläufigen Höhepunkt.

Wenigstens seit dem sechzehnten Jahrhundert gibt es Versuche, die Sprachvielfalt (zunächst Europas) auf ähnliche Weise taxonomisch zu ordnen. So identifizierte Justus Scaliger (1540-1609) u.a. die romanische, griechische, germanische und die slawische Sprachgruppe. Sebastian Münster (1488-1552) erkannte schon 1544 die Verwandtschaft zwischen Ungarisch, Finnisch und Samisch, die wir heute zu den finno-ugrischen Sprachen zählen. Claudius Salmasius (1588-1653) identifizierte Ähnlichkeiten zwischen Latein und Griechisch einerseits und iranischen und indischen Sprachen andererseits. Georg Stiernhelm (1598-1672) erweiterte diese Sprachgruppe um die germanischen, romanischen, slawischen und keltischen Sprachen. Im neunzehnten Jahrhundert entwickelte sich aus diesen vereinzelten Beobachtungen ein systematisches Forschungsprogramm, die historisch-komparative Linguistik. Die Darstellung von Sprachverwandtschaften in einem Stammbaum wird üblicherweise August Schleicher (1821-1861) zugeschrieben, findet sich aber auch schon bei Friedrich von Schlegel (1772-1829).

Wenn eine Gruppe von Sprachen aufgrund ihrer Ähnlichkeit in eine taxonomische Gruppe zusammengefasst wird, liegt die Annahme nahe, dass sie sich aus einer gemeinsamen Ursprache entwickelt haben. So erklärt sich die Verwandtschaft der romanischen Sprachen aus ihrer gemeinsamen Herkunft vom Latein. Für die meisten Sprachfamilien ist diese Ursprache nicht dokumentiert. Dennoch ist die Annahme einer taxonomischen Einheit bei der Sprachklassifikation immer auch eine Hypothese über eine vergangene Sprachstufe. Eine berühmte Formulierung dieser Einsicht stammt von William Jones (1746-1794):

“The Sanscrit language, whatever be its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either, yet bearing to both of them a stronger affinity, both in the roots of verbs and the forms of grammar, than could possibly have been produced by accident; so strong indeed, that no philologer could examine them all three, without believing them ***to have sprung from some common source, which, perhaps, no longer exists****: there is a similar reason, though not quite so forcible, for supposing that both the Gothick and the Celtick, though blended with a*

very different idiom, had the same origin with the Sanscrit, and the old Persian might be added to this family, if this were the place for discussing any question concerning the antiquities of Persia."

(zitiert aus Campbell & Poser 2008:5, Hervorhebung von mir)

Mit Darwins *Entstehung der Arten* (veröffentlicht 1859) setzte sich auch in der Biologie die Einsicht durch, dass taxonomische Klassifikationen historisch zu interpretieren sind, dass also Knoten in einem Stammbaum früheren Arten entsprechen, aus denen die modernen Arten durch Evolution hervorgegangen sind. Fossilienfunde ermöglichen es, derartige Hypothesen empirisch zu überprüfen – analog zu historischen Schriftdokumenten, die Evidenz für frühere Sprachstufen liefern.

Gegen Ende des neunzehnten Jahrhunderts hatte sich somit in Biologie und Linguistik eine bemerkenswert ähnliche Interpretation von Taxonomien durchgesetzt. Der bereits erwähnte Linguist August Schleicher stellte in einem offenen Brief an den Biologen Ernst Häckel (der ihn auf Darwins Arbeit aufmerksam gemacht hatte), fest:

„Was nun zunächst die von Darwin behauptete Veränderungsfähigkeit der Arten im Verlaufe der Zeit betrifft, durch welche, wenn sie nicht bei allen Individuen in gleichem Maasse und in gleicher Weise hervortritt, aus einer Form mehrere Formen hervorgehen (ein Prozess der sich natürlich abermals und abermals wiederholt), so ist sie für die sprachlichen Organismen längst allgemein angenommen. Diejenigen Sprachen, die wir, wenn wir uns der Ausdrucksweise der Botaniker und Zoologen bedienten, als Arten einer Gattung bezeichnen würden, gelten uns als Töchter einer gemeinsamen Grundsprache, aus welcher sie durch allmähliche Veränderung hervorgingen. Von Sprachsippen, die uns genau bekannt sind, stellen wir eben so Stammbäume auf, wie diess Darwin (S. 121) für die Arten von Pflanzen und Thieren versucht hat."

(Schleicher 1863: 14f.)

2. Die komparative Methode in der Linguistik

Das zentrale Kriterium, um die Verwandtschaft zweier Sprachen zu demonstrieren, sind Ähnlichkeiten im Wortschatz. Man vergleiche etwa die Wörter für *eins, zwei, ich, wir,* und *Vater* in einer Stichprobe europäischer Sprachen in Tabelle 1.

Deutsch	eins	zwei	ich	wir	Vater
Englisch	one	two	I	we	father
Niederländisch	een	twee	ik	wij	vader
Schwedisch	en	två	jag	vi	far
Spanisch	uno	dos	yo	nostros	padre
Italienisch	uno	due	io	noi	padre
Rumänisch	unu	doi	eu	noi	tată
Polnisch	jeden	dwa	ja	my	ojciec
Tschechisch	jedna	dva	já	my	otec
Kroatisch	jedan	dva	ja	mi	otac
Ungarisch	egy	kettő	én	mi	apa

Tabelle 1: Multilateraler lexikalischer Vergleich

Die germanischen Sprachen Deutsch, Englisch, Niederländisch und Schwedisch bilden eine natürliche Gruppe, da sie z.B. für *eins* einsilbige Wörter mit einem *n* verwenden, für *zwei* einsilbige Wörter mit einem *t*-Laut, gefolgt von einem *w*-Laut am Anfang, für *Vater* Wörter mit der Lautfolge *fa*- o.ä. am Anfang usw. Gleichermaßen liegt es nahe, die romanischen Sprachen Spanisch, Italienisch und Rumänisch zusammenzufassen (Lautfolge *un*- im jeweiligen Wort für *eins*, initiales *d* im Wort für *zwei*, initiales *no*- für *wir* usw.), wie auch die slawischen Sprachen Polnisch, Tschechisch und Kroatisch (*jed*- für *eins*, *ja* für *ich* usw.). Das ungarische Wort für *wir* entspricht zwar den entsprechenden slawischen Wörtern, davon abgesehen gibt es jedoch keine auffälligen Ähnlichkeiten.

Die Tabelle zeigt auch Gemeinsamkeiten zwischen allen aufgeführten Sprachen außer dem Ungarischen. So ähneln sich etwa die romanischen und slawischen Wörter für *zwei*. Auch die entsprechenden germanischen Wörter sind ähnlich; dem initialen *d*- im romanischen und slawischen entsprechen aber einem germanischen *t*- (mit der Ausnahme des Deutschen, das hier ein *z*- hat). Ein systematischer Vergleich eines größeren Wortschatzes zeigt, dass die Korrespondenz zwischen germanischem *t* und einem *d* in anderen indoeuropäischen Sprachen systematisch ist: vgl. Lateinisch ***d****ecem*, Griechisch ***d****eka*, Kroatisch ***d****esat* vs. Englisch ***t****en*, Niederländisch ***t****ien*, Schwedisch ***t****io*, oder Italienisch *pie****d****e*, Litauisch *pė****d****a*, Sanskrit *pā****d****a* vs. Engisch *foo****t***, Schwedisch *fo****t***, Niederländisch *voe****t***. Gleichermaßen entspricht ein germanisches *f* (wie in Englisch ***f****ather*) nicht nur in romanischen, sondern auch in den anderen indoeuropäischen Sprachen einem *p* (wie z.B. durch den Anlaut in den schon zitierten Übersetzungen von *Fuß* illustriert).

Aus einer Fülle derartiger Beobachtungen zogen historische Linguisten im späten neunzehnten Jahrhundert folgende Schlussfolgerungen:

- Sprachwandel geht mit systematischem Lautwandel einher.
- Ein Lautwandel-Prozess (ein sogenanntes „Lautgesetz“, wie z.B. die Ersetzung von *p* durch *f*) betrifft ausnahmslos alle Wörter einer Sprache.

Systematische Lautkorrespondenzen zwischen gleichbedeutenden Wörtern verschiedener Sprachen sind demnach Evidenz dafür, dass diese Wörter **Kognaten** sind, also vom selben Wort der gemeinsamen Ursprache der verglichenen Sprachgruppe abstammen. Betrachten wir nochmals die genannten Wörter für *Fuß*. Da sich die Lautfolge *f-t* nur in germanischen Sprachen findet, die korrespondierende Lautfolge *p-d* jedoch in Sprachen aus verschiedenen Sprachgruppen (u.a Romanisch, Baltisch, Indisch), muss *p-d* dem Lautstand in der gemeinsamen Ursprache entsprechen. Im Urgermanischen fand ein systematischer Lautwandel von *p* nach *f* und von *d* nach *t* statt (die sogenannte „Germanische Lautverschiebung“; auch „Grimmsches Gesetz“).

Wenn ein Lautgesetz identifiziert ist, kann man es sozusagen rückwärts anwenden und die Urform rekonstruieren, aus der die attestierten Kognaten abgeleitet sind, auch wenn es keine schriftlichen Zeugnisse der Ursprache gibt. Das Urindoeuropäische Wort für Fuß z.B. war vermutlich *ped* oder *pod.*

Die Kenntnis der Lautgesetze erlaubt eine zuverlässigere Identifizierung von Kognaten, als es rein durch Augenschein möglich ist. So liegt etwa die Vermutung nahe, dass Lateinisch *caput* (gesprochen *kaput*; ‚Kopf’) mit dem Deutschen *Kopf* kognat ist. Wir wissen aber, dass ein *k* im Lateinischen und anderen indoeuropäischen Sprachen systematisch einem germanischen *h* entspricht (vgl. Lateinisch ***c****entum* – gesprochen ***k****entum* –, Griechisch *he****k****aton*, Bretonisch ***k****ant* vs. Englisch ***h****undred*, Norwegisch ***h****undre*, Isländisch ***h****undrað*). Der deutsche Kognat zu *caput* ist in der Tat *Haupt* (vgl. Englisch *head* etc.), während *Kopf* ursprünglich ‚Tasse’ oder ‚Schüssel’ bedeutet (vgl. *Kappe* oder Englisch *cup*) und trotz der augenfälligen Ähnlichkeit nicht mit *caput* verwandt ist.

Lautgesetze erlauben es auch, Kognaten zu identifizieren, die sich augenscheinlich nicht ähneln. So sind etwa das deutsche *Hundert* und das russische *sto* kognat: Das urindoeuropäische Wort für ‚Hundert’ lässt sich als *kmtom* rekonstruieren. In der Geschichte des Deutschen wurde das *k* (über Zwischenstufen) durch *h* und das *t* durch *d* ersetzt, während sich in der Geschichte des Russischen das *k* zu einem *s* wandelte.

Um die Verwandtschaft zweier Sprachen zu demonstrieren, ist es nach den Standards der historischen Linguistik unerlässlich, die Existenz einer hinreichend großen Anzahl von Kognaten zu demonstrieren. Diese Aufgabe wird dadurch erschwert, dass Sprachen häufig Wörter aus anderen – meist benachbarten – Sprachen **entlehnen**. Häufig betrifft das technische und kulturelle Innovationen, die gemeinsam mit dem entsprechenden Konzept aus einer anderen Sprachgemeinschaft übernommen werden. In der Deutschen Sprache finden sich z.B. im Altertum und Mittelalter lateinische Entlehnungen (lat. *fructus* > dt. *Frucht*, lat. *tegula* > dt. *Ziegel*, lat. *claustrum* > dt. *Kloster*), in der frühen Neuzeit aus dem Italienischen (it. *banca* > dt. *Bank*, it. *capitale* > dt. *Kapital*) und in der jüngsten Gegenwart aus dem Englischen (*Computer, chillen, chatten*). Der Grundwortschatz einer Sprache, also Wörter für alltägliche, nicht kulturabhängige Alltagskonzepte (z.B. einfache Zahlwörter, Körperteil- und Verwandtschaftsbezeichnungen usw.) sowie grammatische Morpheme wie Pronomina, Hilfsverben oder Artikel, ist weniger häufig von Entlehnungen betroffen. Daher sind Kognaten v.a. dann Evidenz für Sprachverwandtschaft, wenn sie dem Grundwortschatz angehören. Allerdings kommen auch in diesem Bereich Entlehnungen vor. So ist das englische Pronomen *they* eine Entlehnung aus dem Altnordischen *þeir*, und das Finnische *tytär* ‘Tochter’ ist eine baltische Entlehnung (vgl. Estnisch *tütar*).

Gemeinsamkeiten in der Grammatik zweier Sprachen geben ebenfalls Hinweise auf Sprachverwandtschaft. Man vergleiche etwa das Konjugationsparadigma des Hilfsverbs *sein* im Präsens im Lateinischen und im Sanskrit:

	Latein	**Sanskrit**
(ich) bin	*sum*	*asmi*
(du) bist	*es*	*asi*
(es) ist	*est*	*asti*
(wir) sind	*sumus*	*smas*
(ihr) seid	*estis*	*stha*
(sie) sind	*sunt*	*santi*

Besonders starke Argumente für Sprachverwandtschaft sind Unregelmäßigkeiten in grammatischen Paradigmen, die in beiden Sprachen vorkommen, wie z.B. Deutsch *gut – besser – am Besten* vs. Englisch *good – better – best*.

3. Etablierte Sprachfamilien

Anhand dieser Kriterien lassen sich die ca. 6 000 lebenden Sprachen in ca. 150 Sprachfamilien gruppieren. Nicht erfasst sind dabei die ca. 120 isolierten Sprachen, für die

sich keine Verwandtschaft zu einer anderen lebenden Sprache nachweisen lässt. Dazu gehören u.a. das Baskische und das Koreanische.

Die gegenwärtig am besten untersuchte Sprachfamilie ist das Indoeuropäische (teilweise auch „Indogermanisch“). Wie oben bereits ausgeführt, wurde die Verwandtschaft des Sanskrit und anderer indischer Sprachen mit dem Griechischen und Lateinischen, aber auch mit den germanischen, slawischen und keltischen Sprachen seit Jahrhunderten vermutet. Sie gilt seit dem neunzehnten Jahrhundert als gesicherte Tatsache. Die indoeuropäische Sprachfamilie gliedert sich v.a. in die balto-slawische, die germanischen, die indo-iranische, die keltische und die romanische Sprachgruppe. Weiterhin umfasst sie das Albanische, das Armenische und das Griechische sowie eine Reihe von ausgestorbenen, aber schriftlich dokumentierten Sprachen wie das Hethitische und das Tocharische. Insgesamt umfasst diese Sprachfamilie über zweihundert lebende Sprachen.

Für die indoeuropäische Ursprache konnten mehrere hundert Wörter rekonstruiert werden, für die es in den Tochtersprachen Kognate gibt. Außerdem haben wir eine gute Vorstellung von der Laut- und Formenlehre dieser Sprachstufe.

Es ist immer noch kontrovers, wann diese Sprache gesprochen wurde. Nach gegenwärtigem Kenntnisstand war das zwischen ca. 7 000 v.Chr. und 3 700 v.Chr.

Bei der Etablierung der indoeuropäischen Sprachfamilie und der Rekonstruktion ihrer Ur-Sprache spielten schriftliche Dokumente älterer Sprachstufen eine wichtige Rolle. Die ältesten derartigen Zeugnisse – aus dem Hethitischen – gehen bis ins zweite vorchristliche Jahrtausend zurück. Lediglich die semitischen Sprachen sind vergleichbar gut in älteren Sprachstufen dokumentiert. Für die meisten Sprachfamilien liegen gar keine schriftlichen Zeugnisse vor. Daher ist der Kenntnisstand über das Indoeuropäische eine Art *benchmark* für die Reichweite der komparativen Methode.

Außer dem Indoeuropäischen gibt es in Europa und Asien noch ungefähr 20 Sprachfamilien und ungefähr 10 isolierte Sprachen. (Über die genauen Zahlen differieren die Expertenmeinungen etwas.) Die größten eurasischen Sprachfamilien sind (neben Indoeuropäisch) Sino-Tibetisch (umfasst u.a. chinesische und tibetische Sprachen), Afro-Asiatisch (umfasst u.a. die semitischen, kuschitischen, tschadischen und die Berber-Sprachen), die Turksprachen (die von manchen Experten mit den mongolischen und tungusischen Sprachen zu der größeren Einheit der altaischen Sprachen zusammengefasst werden; diese Klassifikation ist jedoch kontrovers), Tai-Kadai, Austroasiatisch (beide in Südostasien beheimatet), Dravidisch (Südindien) und Uralisch (umfasst u.a. die finno-ugrischen Sprachen wie Ungarisch und Finnisch).

Die afrikanischen Sprachen werden gemeinhin in vier Familien klassifiziert (wobei diese Einteilung nicht unumstritten ist): das bereits erwähnte Afro-Asiatisch, Niger-Kongo (umfasst u.a. die Bantu-Sprachen), Nilo-Saharisch und Khoisan.

Die ca. eintausend lebenden Sprachen der amerikanischen Ureinwohner lassen sich in 50-60 etablierte Sprachfamilien einteilen. Die größten darunter sind die Maya-Sprachen, Oto-Mangue (hauptsächlich Mexiko), Uto-Atztekisch (Mexiko und Südwesten der USA) und Arawakisch (nördliches Südamerika).

Die ca. einhundertfünzig lebenden Sprachen der australischen Ureinwohner fallen in ungefähr ein Dutzend etablierte Sprachfamilien. Die größte darunter ist das Pama-Nyungan.

Die – gemessen an der Anzahl der Sprachen – größte Sprachfamilie ist das Austronesisch. Sie hat eine immense geographische Ausdehnung, von Madagaskar im Westen über Indonesien, Taiwan und Neuseeland bis hin zu Hawaii und der Osterinsel im Osten. Zu den gemessen an der Sprecherzahl größten austronesischen Sprachen gehören Malaiisch, Javanisch, Tagalog und Indonesisch.

Die Region mit der weltweit höchsten Sprachenvielfalt ist Papua-Neuguinea. Bei einer Bevölkerung von nur 6,7 Millionen Einwohnern werden dort – auf einem Gebiet, das kleiner ist als z.B. Spanien – über siebenhundert Sprachen gesprochen. Ungefähr zweihundert davon sind austronesische Sprachen. Die Übrigen gliedern sich in ca. fünfzig etablierte Familie (neben ca. zwanzig isolierten Sprachen), die keine nachweisbaren Beziehungen untereinander oder zu Sprachen auf anderen Inseln oder Kontinenten haben.

4. Grenzen der komparativen Methode

Je weiter die Rekonstruktion von Sprachverwandtschaft in der Zeit zurück reicht, umso spärlicher wird die verfügbare Evidenz. Dafür gibt es mehrere Gründe. Zum einen nimmt die lautliche Ähnlichkeit zwischen Kognaten in verschiedenen Zweigen einer Sprachfamilie mit der Zeit ab. Erinnert sei an das Kognat-Paar *hundert* (deutsch) vs. *sto* (Russisch). Die Verwandtschaft dieser beiden Formen konnte nur etabliert werden, weil Evidenz aus vielen weiteren indoeuropäischen Sprachen vorliegt, die es erlaubt, die entsprechenden Lautwandelprozesse in der Geschichte der beiden Sprachen zu rekonstruieren. Wenn man z.B. die Frage untersucht, ob Indoeuropäisch und Uralisch miteinander verwandt sind, muss man nach der Logik der komparativen Methode Kognaten zwischen dem rekonstruierten Wortschatz des Ur-Indoeuropäischen und des Ur-Uralischen identifizieren. Bei einem Wortpaar wie *hundert/sto* gäbe es keine Anhaltspunkte für den Kognatenstatus.

Weiterhin verändern Worte im Laufe der Sprachgeschichte ihre Bedeutung. Wie oben bereits erwähnt ist das deutsche *Kopf* kognat mit dem englischen *cup*. Da es zwischen den beiden Bedeutungen nur eine sehr lose Beziehung gibt (*Kopf* in der ursprünglichen Bedeutung *Schüssel* ist eine Metapher, die auf der Form des Schädels basiert), lässt sich rein auf der Basis der Fakten über die deutsche und englische Gegenwartssprache nicht ohne weiteres sehen, dass es sich um Kognaten handelt. In diesem Beispiel haben wir zusätzliche Informationen über frühere Sprachstufen sowie über Dialekte des Deutschen, in denen *Kopf* noch die ursprüngliche Bedeutung hat.[1] Derartige Evidenz liegt aber bei der Rekonstruktion weit zurückliegender Sprachstufen nicht vor. Bei der Identifikation von Kognaten muss man also die Möglichkeit semantischen Wandels in Betracht ziehen. Üblicherweise wird für die Identifikation von Kognaten nur verlangt, dass ihre Bedeutung hinreichend ähnlich ist, so dass sich die Differenz durch semantischen Wandel erklären lässt. Wenn man dabei zu großzügig vorgeht, besteht allerdings die Gefahr, dass Wörter, die sich zufällig ähneln, fälschlich als Kognaten betrachtet werden.

Um diesen Effekt zu demonstrieren, habe ich ein kleines Experiment durchgeführt. Am Leipziger Max-Planck-Institut für Evolutionäre Anthropologie wurde eine elektronische Datenbank aufgebaut, die einen Grundwortschatz von vierzig Wörtern aus über fünftausend Sprachen, also der großen Mehrheit der lebenden Sprachen, umfasst. Ich verglich systematisch alle Paare von Sprachen von verschiedenen Kontinenten, also Paare von Sprachen, die mit hoher Sicherheit nicht miteinander verwandt sind (bzw. deren gemeinsame Ur-Sprache Jahrzehntausende zurückliegt, so dass Kognaten nicht mehr nachweisbar sind). Dabei betrachtete ich jeweils die Worte für ‚Baum', ‚Horn' ‚Knochen' und ‚Zahn', also Konzepte, zwischen denen eine gewisse semantische Ähnlichkeit besteht.

Bei ungefähr 2 Prozent dieser Sprachpaare ähneln sich die Worte für wenigstens eines dieser Konzepte so stark, dass man vermuten könnte, es handelt sich um Kognaten. (Zum Beispiel ist das Wort für ‚Blatt' in der amerikanischen Sprache Tilquiapan *bldag*, und das Wort für ‚Zahn' in der ebenfalls amerikanischen Sprache Yagua ist *zahanda*, könnte also *prima facie* mit *Zahn* verwandt sein. Das Risiko, bei fünf Kandidaten eine Zufallsähnlichkeit zu finden, ist mit 2 Prozent zwar nicht zu vernachlässigen, aber eher gering. Wenn man jedoch auch Lautähnlichkeiten zwischen Wörtern mit verwandter, aber verschiedener Bedeutung in Betracht zieht, ändert sich das Bild. Die Übersetzung von ‚Knochen' in der Indianersprache Galice ist z.B. *zan*, also fast identisch mit dem deutschen *Zahn*. In ungefähr 10 Prozent der betrachteten Sprachpaare fand sich wenigstens eine Lautähnlichkeit zwischen je einem der Wörter für die genannten Konzepte. Bei einer generösen Auslegung

[1] In meinem südthüringischen Heimatdialekt etwa bezeichnet ein Köpfchen eine große Kaffeetasse.

von „Lautähnlichkeit“ und „Bedeutungsähnlichkeit“ ist die Wahrscheinlichkeit von Zufallstreffern also sehr hoch. Ein überzeugendes Argument für genuine Sprachverwandtschaft bedarf also vieler derartige lautlich und semantisch ähnlicher Wortpaare, um die Wahrscheinlichkeit von Zufallsähnlichkeit gering zu halten.

Die Identifizierung von Kognaten wird durch Entlehnungen weiter erschwert. Im Wesentlichen gibt es vier Kriterien, um Lehnwörter von ererbten Kognaten zu unterscheiden. Der Grad der Verwandtschaft zweier Sprachen wird durch eine Gesamtauswertung aller relevanter Faktoren abgeschätzt – Anzahl der Kognaten, Ähnlichkeiten in der Grammatik, Grad der Lautänderungen usw. und – so vorhanden – extralinguistische historische Erkenntnisse. Wenn bei einem synonymen Wortpaar eine deutlich größere Ähnlichkeit besteht, als es vom Gesamtbild her zu erwarten wäre, liegt die Möglichkeit einer Entlehnung nahe. Um ein extremes Beispiel zu wählen: Das deutsche Wort *Känguruh* ähnelt dem Wort *gangurru* der australischen Sprache Guugu Yimidhirr, das ebenfalls auf eine Art Känguruh referiert. Da eine genetische Verwandtschaft zwischen Deutsch und Guugu Yimidhirr ausgeschlossen werden kann, muss es sich um eine Entlehnung handeln.

Weiterhin muss die Annahme einer Entlehnung historisch plausibel sein. Die oben erwähnte Annahme, dass das finnische Wort für ‚Tochter’, *tytär*, aus dem Baltischen entlehnt wurde, erfüllt dieses Kriterium, da Finnisch zu den baltischen Sprachen geographisch benachbart ist.

Drittens ist es typisch für Entlehnungen, dass sie in den Sprachen, die mit der Zielsprache verwandt sind, keine Kognaten haben. Das finnische *tytär* etwa hat kein ungarisches Kognat (das ungarische Wort für ‚Tochter’ ist *lány*), obwohl Finnisch und Ungarisch eng verwandt sind.

Typisch für Entlehnungen zwischen genetisch verwandten Sprachen ist schließlich, dass sie nicht den für dieses Sprachpaar charakteristischen Lautkorrespondenzen folgen. Das deutsche Wort *Matjes* etwa ist eine Entlehnung aus dem Niederländischen. Das lässt sich u.a. daran erkennen, dass dem niederländischen *t* auch ein deutsches *t* entspricht. Bei ererbten Kognaten entspricht ein niederländisches *t* einem deutschen *s* (wie in ndl. *wa**t**er* vs. dt. *Wa**ss**er*) oder einem deutschen *z* (wie in ndl. ***t**ijd* vs. dt. ***Z**eit*).

Bei weit zurückliegenden Sprachstufen ist keines dieser Kriterien verlässlich anwendbar. Die Größe des rekonstruierbaren Wortschatzes nimmt mit der Zeittiefe ab, womit es schwerer wird, genuine Lautkorrespondenzen und Kognate zu identifizieren. Auch gibt es bei „tiefen“ Rekonstruktionen gemeinhin keine Gruppe von verwandten Sprachen zum Vergleich, und es liegen keine verlässlichen extralinguistischen historischen Informationen vor.

Aus all den genannten Gründen ist die historische Reichweite der komparativen Methode zum Nachweis von Sprachverwandtschaften begrenzt. Wie bereits erwähnt, liegt der Ursprung der am besten untersuchten Sprachfamilie, des Indoeuropäischen, vermutlich maximal neuntausend Jahre zurück. Es wird gemeinhin angenommen, dass es grundsätzlich unmöglich ist, mit dieser Methode verlässlich Sprachverwandtschaften nachzuweisen, die mehr als zehntausend Jahre in die Vergangenheit zurückreichen.

5. Tiefe Sprachverwandtschaft

In den vergangenen hundert Jahren gab es immer wieder Versuche, diese Schallmauer von zehntausend Jahren zu durchbrechen. Der vermutlich bekannteste derartige Vorschlag betrifft eine mögliche Makro-Familie namens „Nostratisch“, die neben Indoeuropäisch eine Reihe weiterer europäischer und asiatischer Sprachfamilien umfasst. Zwischen den Vertretern dieser Hypothese herrscht kein völliger Konsens über den Umfang des Nostratischen, aber üblicherweise werden Afroasiatisch, Uralisch, Dravidisch, Altaisch (also die Turksprachen sowie die mongolischen und tungusischen Sprachen) und die Eskimo-Aleutischen Sprachen hinzugezählt. Zum Teil werden auch Japanisch, Koreanisch, Tschuktscho-Kamtschadalisch sowie einige sibirische Sprachen als mögliche nostratische Sprachen betrachtet.

Die nostratische Hypothese wurde erstmals 1903 von dem dänischen Linguisten Holger Pedersen (1867-1953) formuliert. In den sechziger Jahren des vorigen Jahrhunderts wurde v.a. von den Moskauer Linguisten Vladislav Illich-Svitych und Aharon Dolgopolsky intensive Forschung dazu betrieben.

Über die Zeittiefe des Nostratischen lässt sich nur spekulieren, da die Evidenz – sofern man sie akzeptiert – keine verlässlichen Verbindungen zu archäologischen Einsichten zulässt. Schätzungen liegen in der Größenordung von zwölftausend Jahren.

Vertreter der nostratischen Hypothese benutzen die komparative Methodologie, die oben beschrieben wurde. Sie untersuchen also v.a. den rekonstruierten Wortschatz der Ur-Sprachen der beteiligten bekannten Sprachfamilien (Ur-Indoeuropäisch, Ur-Uralisch usw.) und identifizieren mögliche Kognaten und Lautkorrespondenzen, um so den Ur-Nostratischen Wortschatz zu rekonstruieren. Die Mehrzahl der Experten betrachtet die nostratische Hypothese jedoch skeptisch bis ablehnend. Diese Skepsis basiert darauf, dass die Wortähnlichkeiten, die die Nostratiker als Evidenz für Kognatheit betrachten, auch Zufallsähnlichkeiten oder das Resultat von Entlehnung sein können.

Weiterhin wurde verschiedentlich vorgeschlagen, die nordkaukasischen Sprachen, Baskisch, Sino-Tibetisch, diverse sibirische Sprachen und die nordamerikanischen Na-Dené-Sprachen in eine Makro-Familie namens Dene-Kaukasisch zusammenzufassen (siehe z.B. Shevoroshkin 1991). Für das Ur-Kaukasisch-Dene wäre dann eine Zeittiefe von ca. zwanzigtausend Jahren anzusetzen. Dieser Vorschlag findet gemeinhin noch weniger Akzeptanz als die nostratische Hypothese.

Der amerikanische Linguist Joseph Greenberg (1915-2001) entwickelte eine völlig neue Methode – den sogenannten **lexikalischen Massenvergleich** – welche die komparative Methode ersetzen soll. Dabei vergleicht man Sprachen nicht paarweise, sondern betrachtet den Wortschatz einer große Gruppe von Sprachen gleichzeitig. Durch lexikalische Ähnlichkeiten lassen sich die betrachteten Sprachen in Gruppen zusammenfassen, die dann Kandidaten für genetische Einheiten sind. Die Gruppierung der Sprachen in der Tabelle zu Beginn von Abschnitt 2 oben in Germanisch, Romanisch, Slawisch und Ungarisch, und die Gruppierung aller Sprachen außer Ungarisch in eine größere Gruppe, entspricht dieser Methode.

Greenberg entwickelte diese Methode in den fünziger Jahren des letzten Jahrhunderts und wandte sie zunächst auf die Klassifikation der afrikanischen Sprachen an. Sie führte ihn zur Annahme der vier großen afrikanischen Familien Afro-Asiatisch, Niger-Kongo, Nilo-Saharisch und Khoisan (vgl. Greenberg 1966). Diese Gruppierungen werden von der Fachwelt weitgehend akzeptiert.

Für die amerikanischen Sprachen kam Greenberg zu einer Klassifikation von nur drei Makro-Familien: Eskimo-Aleut, Na-Dené (eine Familie, die nordamerikanische Sprachen aus dem pazifischen Nordwesten und dem Südwesten der USA umfasst) sowie Amerind (vgl. Greenberg 1987). Die letztere Gruppe umfasst sämtliche Indianersprachen außer den ca. 30 Na-Dené-Sprachen. Die Amerind-Hypothese wird von den Experten nahezu einhellig als nicht überzeugend erachtet. Zum einen wird aus einer eher konservativen Position heraus kritisiert, dass die Etablierung einer genetischen Einheit nur über die Rekonstruktion der gemeinsamen Ursprache (bzw. zumindest eines substantiellen Wortschatzes derselben) und der Lautgesetze, die von der Ursprache zu den bekannte Sprachstufen führen, möglich ist. Aber auch Linguisten, die der Methode des lexikalischen Massenvergleichs gegenüber grundsätzlich offen sind, wenden ein, dass Greenberg seine Methode nicht klar genug definiert habe um wissenschaftlichen Ansprüchen zu genügen. Seine Argumentation für Amerind nutzt auch relativ vage lautliche Ähnlichkeiten zwischen Wörtern mit semantisch lose verwandten Bedeutungen. Es fehlen aber (a) klare Kriterien, welcher Grad an lautlicher und semantischer Ähnlichkeit für die Stipulation von Kognaten notwendig ist und (b) eine

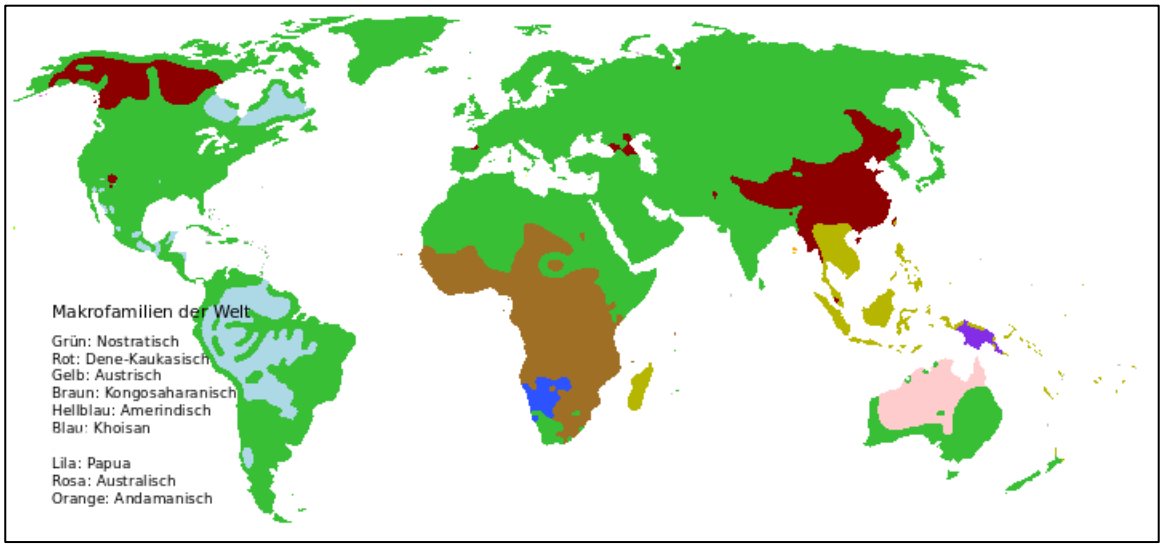

Abbildung 1: Hypothetische Makro-Familien (Graphik aus Wikipedia)

Abschätzung der Wahrscheinlichkeit, dass Ähnlichkeiten auf Zufall oder Entlehnung beruhen.

Dessen ungeachtet gibt es humangenetische Evidenz dafür, dass die Gruppierung der Indianersprachen in Amerind und Na-Dené tatsächlich zwei prähistorischen Einwanderungswellen von Sibirien nach Amerika entspricht. Evidenz dafür wurde in mehreren populationsgenetischen Studien gefunden (vgl. Reich et al. 2012)

Greenberg schlug später weitere Makro-Familien vor, die tiefe Verwandtschaften zwischen etablierten Sprachfamilien widerspiegeln sollen. Der wichtigste derartige Vorschlag betrifft das Euroasiatische (Greenberg 2000/2002). Dieser Vorschlag ist eng mit der nostratischen Hypothese verwandt. Laut Greenberg umfasst das Euroasiatische v.a. Indoeuropäisch, Uralisch, Altaisch, Tschuktscho-Kamtschadalisch, Eskimo-Aleut sowie Japanisch und Koreanisch. Der wichtigste Unterschied zur nostratischen Hypothese besteht darin, dass Greenberg das Afro-Asiatische und das Dravidische nicht zum Euroasiatischen zählt. Greenbergs Argumente für die Annahme des Euroasiatischen basieren hauptsächlich auf lexikalischem Massenvergleich, aber auch auf grammatischen Gemeinsamkeiten zwischen den betrachteten Sprachfamilien.

Merrit Ruhlen, ein Schüler Greenbergs, kombiniert diese und eine Reihe weiterer – im Allgemeinen noch skeptischer beurteilte – Vorschläge für Makro-Familien zu der Annahme, dass sich die Sprachen der Welt in nur acht genetische Einheiten unterteilen (siehe Abb. 1, vgl. Ruhlen 1994).

Ruhlen geht so weit zu behaupten, dass es möglich ist, einen Kernwortschatz des „Proto-Sapiens“ zu rekonstruieren, also der ursprünglichen Sprache der Menschheit. Die generellen methodischen Einwände gegen den lexikalischen Massenvergleich gelten verstärkt auch für Ruhlens Argumentation. Seine „globalen Etymologien“ basieren auf lautlichen und semantischen Ähnlichkeiten zwischen Wörtern aus Sprachen verschiedener Weltregionen. Er gibt aber keine Abschätzung der Wahrscheinlichkeit, dass solche Ähnlichkeiten auch zufällig auftreten können.

6. Bioinformatik

Die Rekonstruktion der Evolutionsgeschichte in der Biologie war bis zur Mitte des zwanzigsten Jahrhunderts methodisch mit der komparativen Methode der historischen Linguistik vergleichbar. Aus phänotypischen Ähnlichkeiten zwischen Organismen entwickeln Biologen eine Taxonomie. Die Knoten in diesem Baum werden mit hypothetischen früheren Entwicklungsstufen identifiziert. Fossilfunde – analog zu schriftlichen Dokumenten früherer Sprachstufen – liefern Evidenz für oder gegen die so gewonnene Rekonstruktion.

Die Entdeckung der molekularbiologischen Funktionsweise der Vererbung veränderte dieses Bild grundlegend. Es ist jetzt gesichert, dass biologische Evolution auf der fehlerhaften Weitergabe von Erbinformation, also auf biomolekularen Mutationen beruht.

Die DNA ist eine Sequenz aus vier Arten von Basen, Adenin (A), Guanin (G), Thymin (T) und Cytosin (C). Bei der Vererbung wird diese Sequenz kopiert und an die Nachkommen weitergegeben. Bei sexueller Fortpflanzung werden die Erbinformationen der Eltern kombiniert, aber jede funktionale Teilsequenz – jedes Gen – entspricht der entsprechenden Sequenz einer der beiden Elternteile.

Mutationen sind Kopierfehler während der Vererbung. Diese können in der Einfügung oder Tilgung einer einzelnen Base bestehen, aber auch in der Ersetzung einer Base durch ein anderes oder Verschiebung ganzer Sequenzen innerhalb des Genoms.

Dreiergruppen von Nukleotid-Basen kodieren jeweils ein Protein. Mutationen führen meistens dazu, dass die DNA-Sequenz in eine andere Protein-Sequenz übersetzt wird. Wenn der resultierende Organismus lebens- und fortpflanzungsfähig ist, wird er die veränderte DNA an seine Nachkommen weitergeben.

Wenn man – vereinfacht – annimmt, dass die Wahrscheinlichkeit einer erfolgreichen Mutation weitgehend konstant ist, lässt sich durch den Vergleich der Erbinformation zweier Organismen die ungefähre Zeit abschätzen, die seit der Aufspaltung ihrer Abstammungslinien vergangen ist.

Abbildung 2: Sequenz-Alinierung

Die Bioinformatik ist eine sehr junge Disziplin an der Schnittstelle zwischen Biologie und Informatik, die sich im Wesentlichen mit dem systematischen Vergleich biomolekularer Sequenzen befasst. Durch moderne Sequenzierungsmethoden wurden von einer Vielzahl von Organismen DNA- und Proteinsequenzen ermittelt. Diese Information ist in elektronischer Form weit zugänglich. Bioinformatiker haben Methoden entwickelt, wie sich ähnliche Teilsequenzen in der Erbinformation mehrerer Organismen ermitteln lassen, sogenannte Alinierungs-Algorithmen.

Betrachten wir zur Illustration ein einfaches Beispiel. Die beiden DNA-Abschnitte ATGCGTCGTT und ATCCGCAT sollen miteinander verglichen werden. Es wird angenommen, dass sie beide durch einfache Mutationen – also durch Einfügen, Tilgen oder Ersetzen einer einzelnen Base – aus der selben Sequenz evolviert sind. Abbildung 2 zeigt zwei Möglichkeiten, die beiden Sequenzen zu alinieren. Beim Alinieren können an beliebigen Stellen in einer der beiden Sequenzen Lücken (angezeigt durch ein Minus-Zeichen) eingefügt werden.

Schwarze Linie verbindet jeweils identische Basen, und rote Linien nicht-identische Basen bzw. Basen mit Lücken. Jede Alinierung stellt eine evolutionäre Hypothese dar. Nicht-Identitäten entsprechen dabei Ersetzungs-Mutationen, und Lücken entweder Tilgungen in der Sequenz mit der Lücke oder Einsetzungen in der anderen Sequenz.

Die Wahrscheinlichkeit der einzelnen Mutationstypen lässt sich abschätzen. Die optimale Alinierung ist diejenige, die dem wahrscheinlichsten evolutionären Szenario entspricht. Die Wahscheinlichkeit eines evolutionären Szenarios wiederum, lässt sich aus der zugehörigen Alinierung abschätzen, indem Lücken und Nicht-Identitäten mit Minuspunkten bewertet werden. Alinierungsalgorithmen finden für ein gegebenes Sequenzpaar und ein Bewertungsschema die optimale Alinierung. Wenn wir für unser Beispiel annehmen, dass sowohl Lücken als auch Nicht-Identitäten mit -1 bewertet werden, wäre die erste Alinierung die optimale Lösung.

	Deutsch	Niederländisch	Englisch	Spanisch	Italienisch
Deutsch	0	1	1,5	4	4
Niederländisch	1	0	1,5	4	4
Englisch	1,5	1,5	0	4	4
Spanisch	4	4	4	0	1,5
Italienisch	4	4	4	1,5	0

Tabelle 2: Geschätzte Zeittiefen

Die Anzahl der Mutationen zwischen zwei Sequenzen gibt eine Abschätzung für die Zeit, die seit der Aufspaltung der beiden Abstammungslinien vergangen ist. Wenn für eine Gruppe von Sequenzen die paarweisen Divergenzzeiten bekannt sind, ist es möglich, den evolutionären Stammbaum für diese Gruppe zu bestimmen.

Das sei wiederum mit einem – erfundenen – Beispiel illustriert. Anstelle von Molekularsequenzen benutze ich hier Sprachen, da die Logik dieselbe ist. Tabelle 2 gibt die geschätzten Divergenzzeiten in Jahrtausenden für eine Gruppe indoeuropäischer Sprachen (die Zahlen sind grobe Schätzungen; das Beispiel dient, wie gesagt, nur zur Illustration). Der Abstand dieses Knotens von der gemeinsamen Ursprache muss dann 4-1,5=2,5 betragen. Weiterhin haben Deutsch und Niederländisch einen geringen Abstand voneinander. Daher werden sie in eine Gruppe zusammengefasst. Der Abstand dieses neuen Knotens im Stammbaum vom gemeinsamen Vorfahr mit Englisch beträgt dann 1,5-1=0,5 und der Abstand vom gemeinsamen Vorfahr mit Spanisch/Französisch 4-1=3. Das Englische hat zum gemeinsamen Vorfahr von Deutsch und Niederländisch dann einen Gesamtabstand von 2,5 und zum gemeinsamen Vorfahr von Spanisch und Italienisch einen Gesamtabstand von 6,5. Daher muss das Englische gemeinsam mit Deutsch und Niederländisch eine Gruppe bilden, deren gemeinsame Ursprache 1,5 Zeiteinheiten zurückliegt. Insgesamt ergibt sich damit der Stammbaum in Abbildung 3.

Die hier informell illustrierte Vorgehensweise entspricht dem „Neighbor Joining“-Algorithmus, der in der Bioinformatik häufig zur Rekonstruktion eines phylogenetischen Baums aus paarweisen Alinierungs-Abständen zwischen Biomolekülen verwendet wird.

7. Anwendung auf Sprachdaten

In den fünfziger Jahren des letzten Jahrhunderts machte der amerikanische Linguist Morris Swadesh (1909-1967) einen Vorschlag, wie Abschätzungen über die historische Distanz zweier Sprachen quantifiziert werden können. Seine Idee basiert auf der Beobachtung, dass Entlehnungen meist kulturspezifisches Vokabular betreffen. Der kulturunabhängige

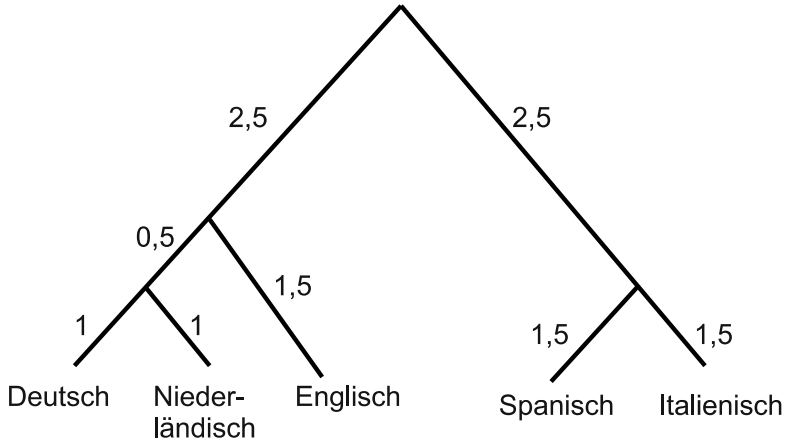

Abbildung 3: Rekonstruktion des Stammbaums anhand der Daten aus Tabelle 1

Grundwortschatz einer Sprache – also Funktionswörter, elementare Verwandtschaftsbezeichnungen, Wörter für Körperteile oder für allgegenwärtige natürliche Objekte wie ‚Wasser', ‚Feuer', ‚Stein' usw. sind vergleichsweise resistent gegen Entlehnungen. Swadesh stellte mehrere Versionen sogenannter „Swadesh-Listen" zur Diskussion, also Listen von derartigen kulturunabhängigen Konzepten (vgl. Swadesh 1971). Die am häufigsten verwendete Version umfasst zweihundert derartige Begriffe. Die Distanz zweier Sprachen kann nach Swadesh durch die Anzahl der Konzepte aus dieser Liste abgeschätzt werden, die in den beiden Sprachen nicht durch kognate Wörter ausgedrückt werden. Swadesh nahm außerdem an, dass die Wahrscheinlichkeit, mit der eine Sprache Elemente des Grundwortschatzes durch nicht-kognate Wörter (durch Entlehnung oder semantischen Wandel) ersetzt, ungefähr konstant ist. Demnach ließe sich durch die so gewonnene Distanz sogar die Zeittiefe des gemeinsamen Ursprungs abschätzen, was die Rekonstruktion eines Stammbaums erlaubt.

Diese Annahmen haben sich als zu optimistisch erwiesen. Die Ersetzungsrate für das Kernvokabular differiert sehr stark sowohl zwischen Sprachen als auch zwischen historischen Phasen. Außerdem ist die Identifikation von Kognaten aus den oben genannten Methoden mit starken Unsicherheiten behaftet.

Gegenwärtig, ein halbes Jahrhundert nach Swadesh, erscheint es jedoch erfolgversprechend, seinen Ansatz wieder aufzugreifen. Die erwähnten bioinformatischen Methoden erlauben eine präzisere Abschätzung der Ähnlichkeiten von Swadesh-Listen, als es manuell möglich ist, und die modernen Methoden der Inferenz phylogenetischer Bäume aus Distanz-Matrizen erlauben es, diese Information besser auszubeuten.

Eine Gruppe um den dänischen Linguisten Søren Wichmann hat in den letzten Jahren eine große Datenbank von Swadesh-Listen zusammengetragen, um eine ausreichende

Abbildung 4: Beispiel-Alinierungen Deutsch/Niederländisch

empirische Basis für die Anwendung moderner computationeller Methoden bereitzustellen (vgl. Wichmann et al. 2011). Die verwendete Liste umfasst lediglich vierzig Konzepte, für welche Übersetzungen in über fünftausend Sprachen gesammelt wurden, also mehr als zwei Drittel aller heute auf der Welt gesprochenen Sprachen. Alle Übersetzungen liegen in einer einheitlichen phonetischen Umschrift vor. Diese Datenbank – das „Automated Similarity Judgment Program“, abgekürzt ASJP, ist frei über das Internet verfügbar.

Im Folgenden werde ich eine Pilotstudie beschreiben, die mit Hilfe bioinformatischer Methoden aus dieser Datenbank Informationen über historische Verwandtschaften zwischen Sprachen extrahiert. Der Ausgangspunkt dafür ist die Idee, dass sich die Distanz zweier Sprachen durch zwei Faktoren abschätzen lässt: 1. die Anzahl der Nicht-Kognaten innerhalb der Swadesh-Liste, und 2. die Anzahl der Lautänderungen, die sich beim Vergleich von Kognaten ergeben.

Der erste Schritt besteht in der paarweisen Alinierung synonymer Wörter aus zwei zu vergleichenden Sprachen. Betrachten wir zur Illustration einen Vergleich Deutsch-Niederländisch. Abbildung 4 zeigt die optimalen Alinierungen für die Paare *ich-ik* und *Laus-luis* (in der im ASJP verwendeten phonetischen Umschrift). Im ersten Paar ist die Distanz 1 und im zweiten Paar 2. Allerdings muss beim Vergleich von Wörtern die Wortlänge in Betracht gezogen werden. In beiden Fällen werden 50% der Laute korrekt aliniert. Im nächsten Schritt wird abgeschätzt, wie wahrscheinlich dieser Grad der Ähnlichkeit bei nicht kognaten Wörtern ist.

Das hängt von den allgemeinen phonetischen Charakteristika der verglichenen Sprachen ab. Werden zwei Sprachen mit einem kleinen und stark überlappenden Lautinventar verglichen, ist die Wahrscheinlichkeit von Zufallstreffern höher als bei zwei Sprachen mit großem und unterschiedlichem Lautinventar. Für den Vergleich Deutsch-Niederländisch ergibt sich beim Alinieren von nicht-synonymen Wörtern eine Trefferwahrscheinlichkeit von ca. 12%. Damit lässt sich, abhängig von der Länge der verglichenen Wörter, abschätzen, wie wahrscheinlich eine Übereinstimmung von 50% ist. Für das Wortpaar *ich-ik* ergibt sich eine Wahrscheinlichkeit von 23%; für *Laus-luis* von nur 7,6%. Das Wortpaar *Laus-luis* ist

	Deutsch	Niederländisch	Englisch	Spanisch	Italienisch
Deutsch	0,00	0,41	0.65	1,00	0,92
Niederländisch	0.41	0,00	0,62	0,97	0,90
Englisch	0.65	0,62	0,00	1,00	0,95
Spanisch	1,00	0,97	1,00	0,00	0,44
Italienisch	0,92	0,90	0,95	0,44	0,00

Tabelle 3: Computationell ermittelte Distanzen

also stärkere Evidenz für die Verwandtschaft von Deutsch und Niederländisch als das Wortpaar *ich-ik*.

Dieser Wert wird für jedes Übersetzungspaar bestimmt und daraus ein Maß der Ähnlichkeit der beiden Sprachen ermittelt. Im letzten Schritt wird abgeschätzt, wie wahrscheinlich ein solcher Ähnlichkeitsgrad unter der Nullhypothese wäre, dass die beiden Sprachen nicht verwandt sind. Für Deutsch-Niederländisch wäre das ein Wert von ca. $7x10^{-193}$. Für ein Paar entfernt verwandter Sprachen wie Deutsch-Hindi ist dieser Wert bei $9x10^{-6}$. Für ein – nach unserem Kenntnisstand – nicht verwandtes Sprachpaar wie Deutsch-Chinesisch liegt der Wert bei 0,82.

Aus diesen Werten wird ein Maß für die Distanz zwischen zwei Sprachen gewonnen. Für die fünf oben betrachteten Beispiel-Sprachen ergeben sich z.B. die Werte in Tabelle 3.

Auch wenn dieses Distanzmaß – besonders für große Distanzen – nicht in einer linearen Beziehung zum zeitlichen Abstand zwischen zwei Sprachen steht, ermittelt der Neighbor-Joining-Algorithmus in den meisten Fällen die korrekten Gruppierungen. Für die betrachteten fünf Sprachen z.B. ergibt sich der korrekte Baum, der in Abbildung 3 dargestellt ist.

In Abbildung 5 ist der so ermittelte Stammbaum für die in der ASJP-Datenbank vertretenen germanischen Sprachen dargestellt. Die wesentlichen Untergruppen werden korrekt wiedergegeben. Es gibt eine primäre Unterteilung zwischen den nordgermanischen (Skandinavien, Island und Faröer) und den westgermanischen Sprachen. Innerhalb der westgermanischen Sprachen bilden die hochdeutschen Dialekte eine eigene Untergruppe. Allerdings gibt es auch signifikante Abweichungen von der etablierten Klassifkation. Englisch z.B. gehört zu den nicht-hochdeutschen westgermanischen Sprachen. Da jedoch Teile seines Kernvokabulars Entlehnungen aus dem Altnordischen, also einer nordgermanischen Sprache sind, wird es den westgermanischen Sprachen nicht eindeutig zugeordent. Auch wird die Aufteilung der hochdeutschen in oberdeutsche (Kimbrisch, Schweizerdeutsch,

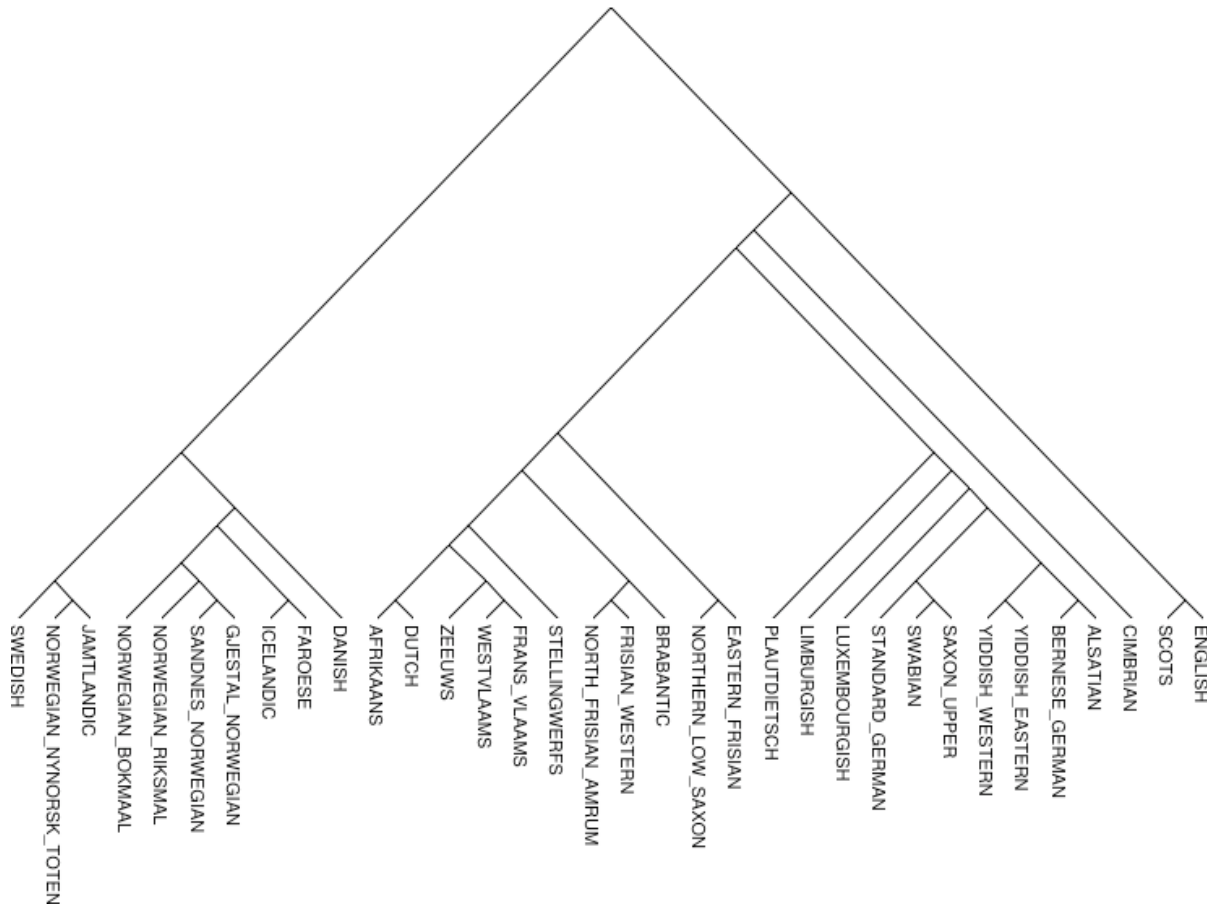

Abbildung 5: Computationell ermittelter Stammbaum der germanischen Sprachen

Elsässisch, Schwäbisch) und mitteldeutsche (Standard-Deutsch, Sächsisch, Luxemburgisch, Limburgisch) Dialekte sowie Jiddisch nicht korrekt erfasst, und Plaudietsch (der niederdeutsche Dialekt der Russland-Mennoniten) wird fälschlich den hochdeutschen Dialekten zugeordnet.

Die so gewonnene automatische Klassifikation ist zwar im Detail fehlerbehaftet, erkennt jedoch die größeren Einheiten mit guter Genauigkeit. Eine automatische Klassifikation der ca. eintausend in ASJP vertretenen eurasiatischen Sprachen ergab einen Baum, durch den von den 73 genetischen Einheiten – Familien wie Indoeuropäisch oder Sino-Tibetisch, und Genera wie Germanisch, Slawisch, Romanisch usw. –, die dem aktuellen Erkenntnisstand der traditionellen Klassifikationsmethode entsprechen (entsprechend dem „World Atlas of Language Structures“; vgl. Haspelmath 2005), 52 Einheiten korrekt erkannt werden, also ca.

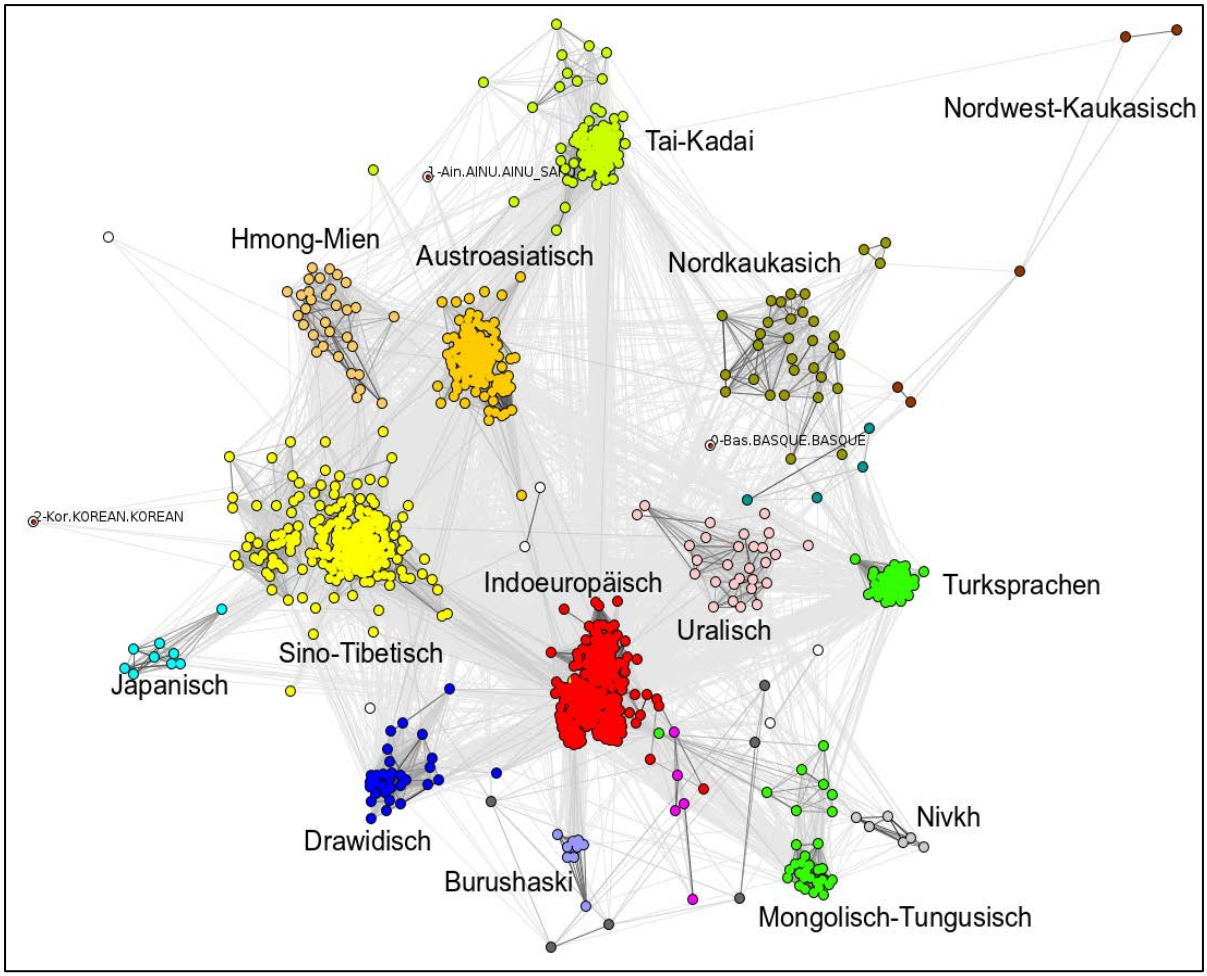

Abbildung 6: Visualisierung der Ähnlichkeiten zwischen den eurasischen Sprachen

71%. Für die afrikanischen Sprachen liegt dieser Wert bei 57%, für die australischen und ozeanischen Sprachen bei 58%, und für die amerikanischen Sprachen bei 74%. Die nicht korrekt erkannten Sprachfamilien werden auch mit guter Näherung erkannt. Z.B. enthält die kleinste automatisch ermittelte Gruppe, die alle 269 vertretenen indoeuropäischen Sprachen umfasst, lediglich 9 nicht-indoeuropäische Sprachen. Die Einheit, welche die 204 Sinotibetischen Sprachen am besten erfasst, enthält fälschlicherweise auch Japanisch und schließt zwei sinotibetische Sprachen aus usw. Besonders interessant ist natürlich die Frage, ob die computergestützte Abschätzung von Sprachverwandtschaft Licht auf die Frage der tiefen Sprachverwandtschaften werfen kann. Aus der Bioinformatik ist jedoch bekannt, dass Methoden wie Neighbor Joining für große Zeittiefen sehr unzuverlässig ist. Es gibt jedoch andere Möglichkeiten der Datenauswertung, die verlässlichere Evidenz liefern.

Am Tübinger Max-Planck-Institut für Entwicklungsbiologie wurde eine Software entwickelt, die es erlaubt, Ähnlichkeiten zwischen Biomolekülen zu visualisieren, ohne dafür einen phylogenetischen Baum anzunehmen (vgl. Frickey/Lupas 2004). Hintergrund dafür ist die Einsicht, dass es zwischen Organismen horizontalen Gen-Transfer gibt, also den Austausch

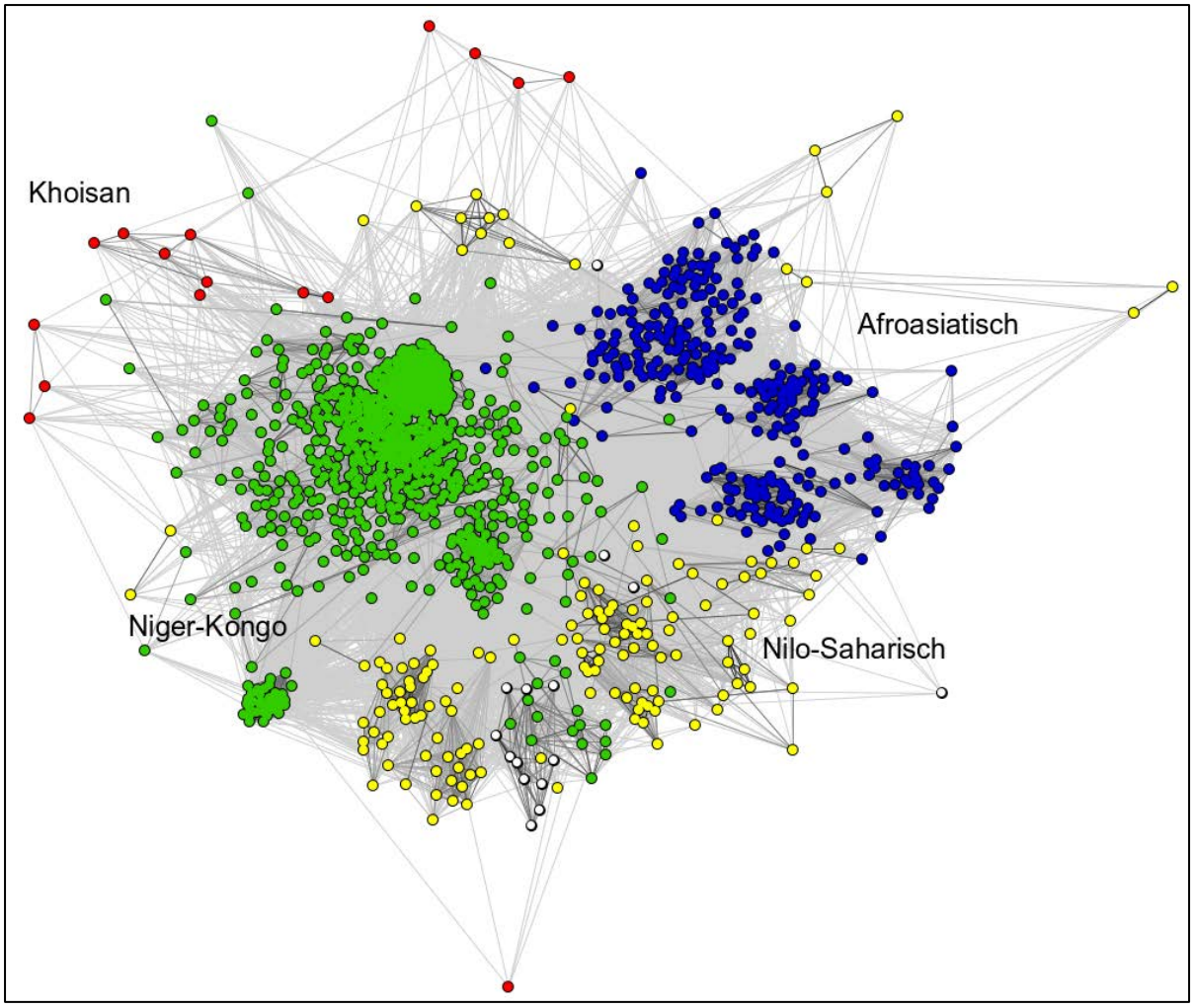

Abbildung 7: Die Sprachen Afrikas

von Erbinformation zwischen nicht verwandten Individuen. Ähnlichkeiten zwischen Gen- oder Proteinsequenzen sind daher nicht immer Indikatoren für gemeinsame Abstammung. Hier besteht eine offensichtliche Parallele zu Entlehnungen zwischen Sprachen. Mit der genannten Software – CLANS – werden schwache Ähnlichkeiten sichtbar gemacht. Sie ist daher gut geeignet, entfernte Verwandtschaften – oder eben Spuren von horizontalem Gentransfer – zu entdecken.

Abbildung 6 zeigt eine Visualierung der Ähnlichkeitenmuster innerhalb der eurasiatischen Sprachen. Die traditionelle Klassifikation ist durch unterschiedliche Farbgebung markiert. Wie leicht zu erkennen ist, bilden die zu einer Famile gehörigen Sprachen jeweils enge Cluster. Eine Ausnahme bilden lediglich die altaischen Sprachen (dargestellt in Grün), die in zwei deutlich getrennte Gruppen zerfallen, die Turksprachen und die Mongolisch-Tungusischen Sprachen. Hierzu ist anzumerken, dass die Zusammenfassung dieser beiden Gruppen zur altaischen Famile auch in der komparativen historischen Linguistik kontrovers ist. Die hier vorgestellte Methode liefert somit weitere Evidenz gegen die altaische

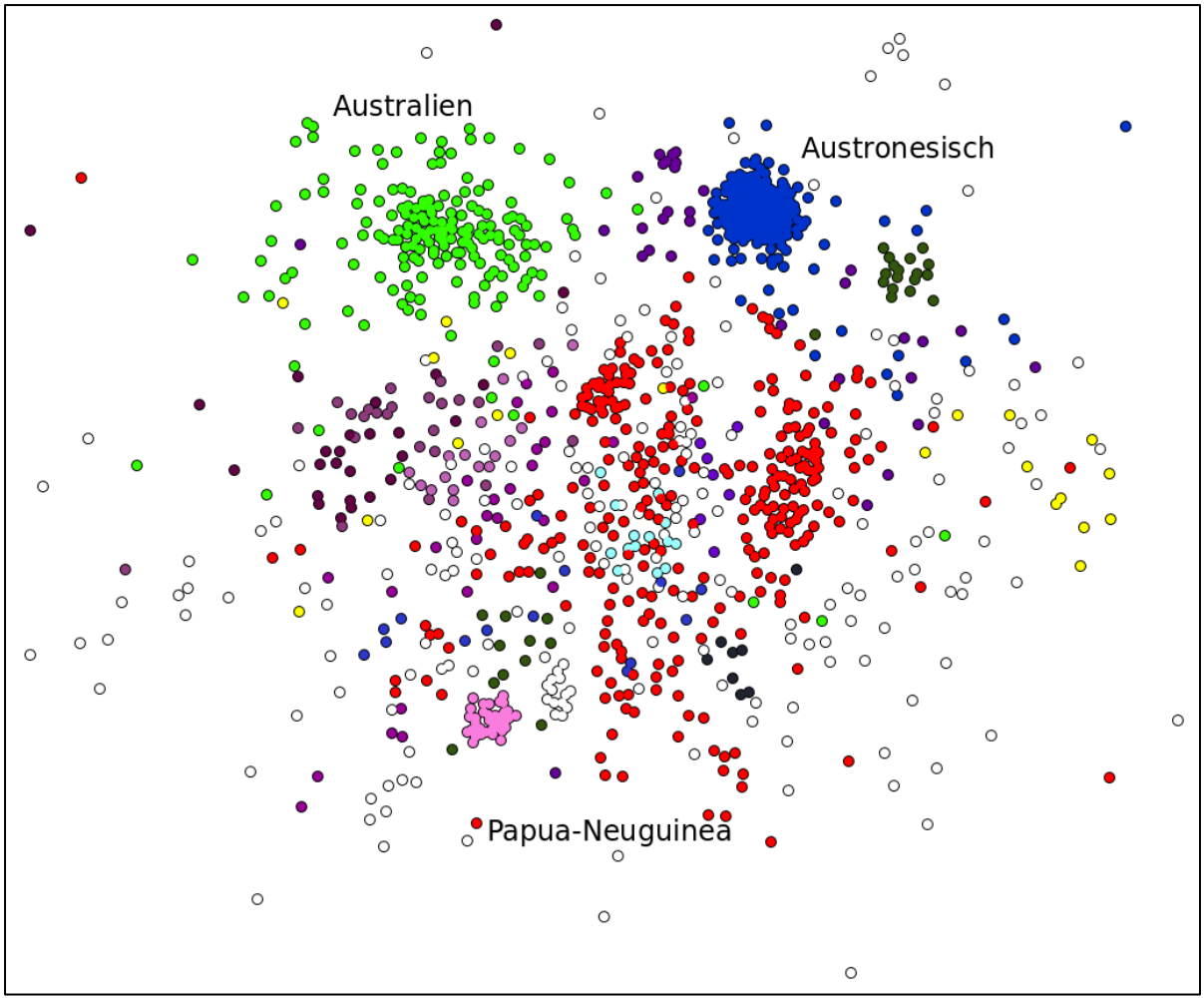

Abbildung 8: Die Sprachen des indo-pazifischen Raums

Hypothese. Bemerkenswert sind jedoch vor allem die globalen Muster. Die Graphik zerfällt, grob gesagt, in zwei Meta-Cluster. Die linke obere Hälfte umfasst die ost- und südost-asiatischen Sprachen, also Sino-Tibetisch, Hmong-Mien, Austroasiatisch, Tai-Kadai sowie Japanisch. In der rechten unteren Hälfte finden sich die nord- und zentralasiatischen sowie die europäischen Sprachen, also Drawidisch, Burushaski (eine im Norden Pakistans beheimatete isolierte Sprache), Indoeuropäisch, Uralisch, Mongolisch-Tungusisch, diverse sibirische Sprachen, die Turksprachen sowie die kaukasischen Sprachen. Diese Gruppe hat eine bemerkenswerte Überlappung mit dem hypothetischen Nostratisch bzw. Euroasiatisch. Allerdings werden Burushaski und die kaukasischen Sprachen nicht zu diesen möglichen Makro-Familien gezählt.

Die afrikanischen Sprachen sind weniger klar gegliedert, wie die Graphik in Abbildung 7 zeigt. Die Niger-Kongo-Sprachen (in grün) und die Afroasiatischen Sprachen (in Blau) bilden zwar erkennbare Cluster, die aber weniger klar artikuliert sind als die euroasiatischen Sprachfamilien. Die Nilo-Saharischen Sprachen (in Gelb) sind nicht als

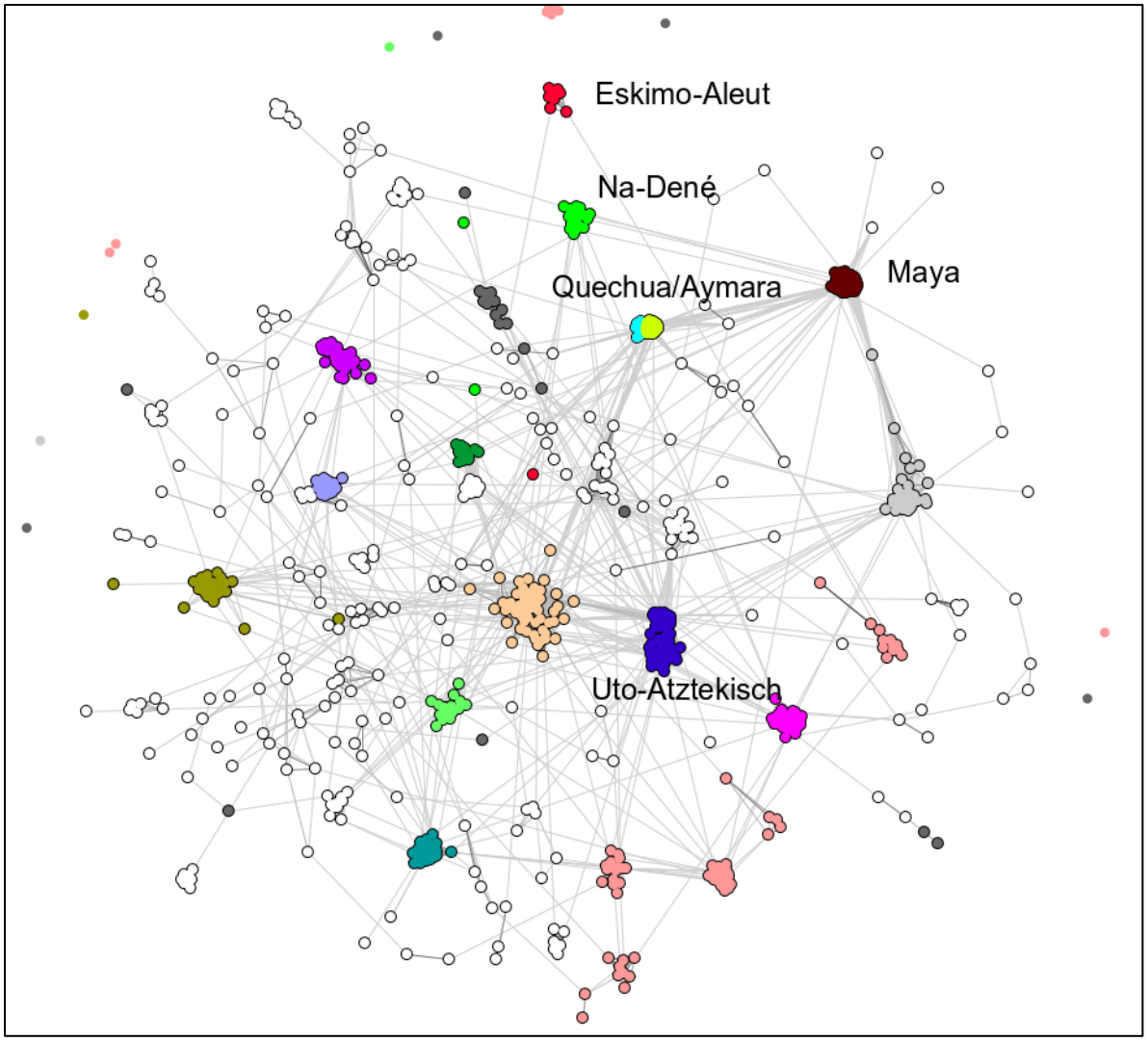

Abbildung 9: Die Sprachen Amerikas

zusammenhängender Cluster erkennbar, und ihre Abgrenzung zu den Niger-Kongo-Sprachen ist nicht sehr scharf. Die Khoisan-Sprachen schließlich (in Rot) zerfallen deutlich in drei Gruppen und sind nicht als Einheit erkennbar.

Die Sprachvielfalt des Indopazifischen Raums, also Australiens sowie der Inselwelt des Indischen und des Stillen Ozeans, ist in Abbildung 8 dargestellt. Sie ist ebenfalls nicht sehr stark gegliedert.

Die australischen Sprachen (in Grün) sowie die Austronesischen Sprachen (in Blau) bilden identifizierbare Cluster. Die verbleibenden Sprachen – fast ausschließlich Papua-Sprachen – lassen sich jedoch nicht einfach in größere Einheiten zusammenfassen.

Die indigenen amerikanischen Sprachen (Abbildung 9) bilden eine Vielzahl von kleinen Clustern, jedoch kein offensichtlichen größeren Gruppierungen. Insbesondere unterscheiden sich die Eskimo-Aleutischen Sprachen (in Rot) und die Na-Dené-Sprachen (in Grün) nicht

sehr stark von den anderen Sprachfamilien. Für Greenbergs Hypothese einer amerindischen Makro-Familie, die alle amerikanischen Sprachen außer Eskimo-Aleut und Na-Dené umfasst, findet sich also keine Evidenz. Bemerkenswert ist, dass die Quechua-Sprachen (Hellgrün) und die geographisch benachbarten Aymara-Sprachen (Türkis) ein kompaktes Cluster bilden. Die starke Affinität zwischen diesen beiden Sprachfamilien ist den Experten seit langem bekannt. Es ist gegenwärtig kontrovers, ob diese Ähnlichkeiten durch gemeinsame Abstammung oder als Resultat jahrhundertelangen Sprachkontakts zu erklären sind.

8. Ausblick

Die Anwendung bioinformatischer Methoden war in den letzten Jahren mehrfach Thema von vielbeachteten Publikationen in *Nature*, *Science* und anderen prominenten Fachzeitschriften. Diese Arbeiten haben heftige Kontroversen ausgelöst. Erwähnenswert ist beispielsweise die Arbeit von Gray & Atkinson (2003). Auf der Basis von Kognatenlisten für 87 indoeuropäischen Sprachen und phylogenetischer Rekonstruktion kommen die Autoren zu dem Ergebnis, dass die indoeuropäische Ursprache vor 9 800–7 800 Jahren gesprochen wurde. Das ist mit der sogenannten „anatolischen Theorie“ des Ursprungs des Indoeuropäischen kompatibel, wonach die Urindoeuropäer anatolische Bauern waren und die Ausbreitung des indoeuropäischen mit der Verbreitung der Landwirtschaft verbunden ist. Die Mehrheitsmeinung der Indoeuropäisten ist jedoch, dass die Urindoeuropäer vor ca. 6 000 Jahren in Südrussland lebten und einen nomadischen Lebensstil pflegten – die sogenannte „Kurgan-Hypothese“. Für die Kurgan-Theorie spricht u.a., dass der rekonstruierte Wortschatz des Urindoeuropäischen z.B. Wörter für Rad, Achse, Deichsel, Nabe usw. umfasst. Da das Rad vermutlich erst vor ca. 6 000 Jahren erfunden wurde, spricht das gegen eine größere Zeittiefe des Indoeuropäischen.

In einem aktuellen Beitrag für *Science* (Bouckaert et al. 2012) nutzen die Autoren phylogeographische Methoden aus der Evolutionsbiologie, um den Prozess der räumlichen Ausdehnung der indoeuropäischen Sprachfamilie zu rekonstruieren. Sie kommen ebenfalls zu dem Ergebnis, dass ein Ursprung vor 9 500 bis 8 000 Jahren in Anatolien das wahrscheinlichste Szenario ist.

Diese Vorschläge wurden von Vertretern der traditionellen historisch-komparativen Linguistik zum Teil heftig kritisiert. Beanstandet wurde u.a., dass sich die beiden genannten Arbeiten nur auf einen Teil der verfügbaren Evidenz stützen – Kognaten im Basisvokabular einer Auswahl aller indoeuropäischer Sprachen –, und dass die Ergebnisse im Detail Einsichten widersprechen, die eigentlich als gesichert gelten.

Diese Debatte ist zweifelsohne der Beginn einer längeren Methodendiskussion, die das Feld der historischen Linguistik grundlegend verändern wird. Die Bioinformatik stellt für die Sprachwissenschaft einen neuartigen Werkzeugkasten bereit, und die Linguisten beginnen in diesen Jahren, die neuen Möglichkeiten auszutesten und auf ihre Tauglichkeit zu überprüfen. Vermutlich wird die direkte Anwendung bioinformatischer Algorithmen auf sprachliche Daten in wenigen Jahrzehnten als naiv erscheinen, und wir werden über Techniken verfügen, welche viel besser auf die spezifischen Eigenschaften des Sprachwandels kalibriert sind. In jedem Fall verspricht die historische Sprachwissenschaft in näherer Zukunft spannend zu werden.

Literatur

Atkinson, Quentin D./Gray, Russell (2005): “Curious parallels and curious connections – phylogenetic thinking in biology and historical linguistics”, in: Systematic Biology 54, S. 513-526.

Bouckaert, Remko/Lemey, Philippe/Dunn, Michael/Greenhill, Simon J./Alekseyenko, Alexander V./Drummond, Alexei J./Gray, Russell D./Suchard, Marc A./Atkinson, Quentin D. (2012): “Mapping the origins and expansion of the Indo-European language family”, in: Science 337, S. 957–960.

Campbell, Lyle/Poser, William J. (2008): Language Classification: History and Method, Cambridge, UK: Cambridge University Press.

Darwin, Charles (1859): On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, London: John Murray.

Darwin, Charles (1871): The descent of man, and selection in relation to sex, London: John Murray.

Frickey, Tancred/Lupas, Andrei N. (2004): “Clans: a Java application for visualizing protein families based on pairwise similarity”, in: Bioinformatics 20, S. 3702-3704.

Gray, Russell. D./Atkinson, Quentin D. (2003): “Language-tree divergence times support the Anatolian theory of Indo-European origin”, in: Nature 426, S. 435-439.

Greenberg, Joseph H. (1966): The languages of Africa, Bloomington: Indiana University Press.

Greenberg, Joseph H. (1971): “The Indo-Pacific hypothesis”, in: Current trends in linguistics 8, S. 809–871.

Greenberg, Joseph H. (1987): Language in the Americas, Stanford: Stanford University Press.

Greenberg, Joseph H. (2000): Indo-European and Its Closest Relatives: Grammar, Stanford: Stanford University Press.

Greenberg, Joseph H. (2002): Indo-European and Its Closest Relatives: Lexicon, Stanford: Stanford University Press.

Haspelmath, Martin (2005): The World Atlas of Language Structures, Oxford: Oxford University Press.

Reich, David/Patterson, Nick/Campbell, Desmond/Tandon, Arti/Mazieres, Stéphane et al. (2012): “Reconstructing Native American population history”, in: Nature 488, S. 370–374.

Ruhlen, Merrit (1994): The Origin of Language: Tracing the Evolution of the Mother Tongue, New York: John Wiley and Sons.

Schleicher, August (1863): Die Darwinsche Theorie und die Sprachwissenschaft: Offenes Sendschreiben an Herrn Dr. Ernst Häckel, ao Prof. der Zoologie u. Director des zool. Museums ad Univ. Jena, Böhlau.

Shevoroshkin, Vitaliy V. (Hrsg.) (1991): Dene-Sino-Caucasian Languages. Bochum: Brockmeyer.

Swadesh, Morris (1971): The Origin and Diversification of Language, Chicago: Aldine.

Wichmann, Søren/Müller, André/Velupillai, Viveka/Wett, Annkatrin/Brown, Cecil H./Molochieva, Zarina/Sauppe, Sebastian/Holman, Eric W./Brown, Pamela/Bishoffberger, Julia/Bakker, Dik/List, Johann-Mattis/Egorov, Dimitry/Belyaev, Oole/Urban, Matthias/Mailhammer, Robert/Geyer, Helen/Beck, David/Korovina, Evgenia/Epps, Pattie/Valenzuela, Pilar/Grant, Anthony/Hammarström, Harald (2011): The ASJP database (version 14). http://email.eva.mpg.de/ wichmann/listss14.zip.