Mathematische Methoden beim Sprachvergleich

Bernhard Ganter
Institut für Algebra
ganter@math.tu-dresden.de

Johann Tischler
Allgemeine und Vergleichende Sprachwissenschaft
tischler@tudurz.urz.tu-dresden.de

Vorlesung im Wintersemester 1998/99
an der Technischen Universität Dresden

Die Unterlagen zur Vorlesung werden im Laufe des Semesters an dieser Stelle nach und nach zugänglich gemacht werden.
Achtung: das Skriptum wird derzeit überarbeitet; es kommt deshalb vor, dass sich im Laufe des Semesters auch solche Seiten noch ändern, die schon hier abrufbar waren.

Inhaltsverzeichnis

1 Einführung **5**
- 1.1 Vorrede 5
- 1.2 Ein spielerisches Experiment 6
- 1.3 Sprachvergleich und Stammbäume 10
- 1.4 Das Stammbaummodell heute 14

2 Eine Formalisierung des Stammbaummodells **15**
- 2.1 Ein synoptisches Axiomensystem 15
  - 2.1.1 Die Annahmen zum qualitativen Baummodell 16
- 2.2 Stimmt das Modell? 18
- 2.3 Diagramme und erste Folgerungen 20
- 2.4 Bäume ohne Wurzel 21

3 Stammbaumrekonstruktion **25**
- 3.1 Stammgruppen 25
- 3.2 Das zweite Kopierexperiment 27
- 3.3 Versus-Relationen und Implikationen 31
- 3.4 Charakterisierung der Versus-Relationen. 34

4 Anhang **37**
- 4.1 Die Anfänge der Vergleichenden Sprachwissenschaft 37

Kapitel 1

Einführung

1.1 Warum wir eine Vorrede brauchen.

Der vorliegende Text entstand aus Vorlesungen, die von einem Sprachwissenschaftler und einem Mathematiker gemeinsam gehalten wurden. Es geht um die Frage, ob die Mathematik brauchbare Hilfsmittel für die Vergleichende Sprachwissenschaft liefern kann.

Die Sprachwissenschaften stellen sich, was ihre Nähe zu mathematischer Methodik angeht, ganz unterschiedlich dar. Während besonders die Computerlinguistik auf vielerlei Formalisierungen angewiesen ist, steht in der Historischen Sprachwissenschaft die inhaltliche Arbeit im Vordergrund. Auf Nachfrage erklären Sprachwissenschaftler gelegentlich, dass sie die Mathematik *nicht verstehen*, was gewöhnlich eine spöttische Umschreibung dafür ist, dass sie sich von mathematischer Methodik *nichts Gutes erwarten*. Sie sind überzeugt, dass sich ihre kenntnisreichen sorgfältigen Analysen nicht durch Automatismen ersetzen lassen, und damit haben sie ja auch völlig recht. Sie wissen zudem aus Erfahrung, dass plump vordergründige Anwendung mathematischer Werkzeuge zu unsinnigen Ergebnissen führt.

Aber auch die Vergleichende Sprachwissenschaft kann sich der Formalisierung gar nicht entziehen. Sie stützt sich auf umfangreiches Datenmaterial, und längst wird dieser Wissensschatz elektronisch verwaltet und bearbeitet, was ganz zwangsläufig eine computergerechte Darstellung erfordert. Will die Sprachwissenschaft Herrin im eigenen Wissenschaftshaus bleiben, dann muss sie sich deshalb mit der Formalisierung ihrer Daten auseinandersetzen. Vielleicht kann diese Vorlesung dafür Impulse geben.

Eine fächerverbindende Vorlesungsreihe wie diese zielt auf Verständigung. Wir benutzen, so gut es geht, die Gemeinsprache und nehmen gern in Kauf, dass der Text dadurch weniger wissenschaftlich erscheint. Natürlich werden wir die einfachen Gedanken in den Vordergrund stellen und die weiterführenden in Literaturhinweise verlagern müssen. Das gilt gleichermaßen für die mathematischen wie für die sprachwissenschaftlichen Argumente.

Fächerverbindene Wissenschaft kann nicht voraussetzen, dass die Beteiligten das jeweils andere Fach zusätzlich lernen müssen. Arbeitsteilung ist notwendig und erwünscht, sie erfordert aber eine sichere Verständigung, eine „Schnittstelle“, über deren Handhabung sich die Partner einig sind. Dieser Einigungsprozess ist wirklich aufregend, und er ist schwierig. Es ist ja die Stelle, wo sich jeder auf die Denkweise des anderen Fachs einlassen muss. Gern tut man das nicht, weil die andere Arbeitsweise ungewohnt und scheinbar sinnwidrig ist. Den Sprachwissenschaftler kann befremden, dass die Mathematiker *über nichts Konkretes* reden. Ihre Theorien beziehen sich immer auf Formalismen und niemals auf die Wirklichkeit! Der Mathe-

matiker ist irritiert, weil die Sprachwissenschaft *nicht genau weiß, wovon sie redet*, denn sie führt ihre Arbeitsbegriffe nicht im Stil mathematischer Definitionen ein. Eine sprachwissenschaftliche Arbeit ohne Bezug auf konkrete Sprachen würde nicht akzeptiert; eine mathematische Publikation ohne zweifelsfreie Definitionen gilt als handwerklich unzulänglich.

Selbst in der Frage, was eigentlich als ein wissenschaftliches Ergebnis anzusehen ist, unterscheiden sich die Fachrichtungen: Einem Mathematiker erscheint es unzureichend, wenn eine sprachwissenschaftliche Forschungsarbeit in eine Analyse mündet, die zwar detailreich ist, aber nicht in greifbare Einzelaussagen zusammengefasst wird. Umgekehrt hat ein Sprachwissenschaftler wenig Verständnis, wenn der Ertrag einer mathematischen Arbeit in ein paar abstrakten Theoremen besteht, die der Autor zudem nicht als Meinungsäußerung verstanden wissen will, sondern darauf besteht, dass sie unumstösslich richtig sind, weil er sie ja *bewiesen* hat.

Natürlich sind beide Denkweisen richtig und stimmig. Es erfordert aber Offenheit, sich darauf einzustellen. Deshalb brauchen wir diese Vorrede. Das Anliegen dieser Vorlesungsreihe ist nicht in erster Linie die Mitteilung mathematischer Methoden. Uns geht es um *Methodenverständnis*, daraus resultierende *Methodensicherheit* und darauf begründete *Methodenakzeptanz*. Wir meinen, dass ein Sprachwissenschaftler mathematische Methodik nur benutzen sollte, wenn er sie sich zu Eigen gemacht hat und selbstbewusst beurteilen kann, welche sprachwissenschaftliche Bedeutung und welchen Sinn die Ergebnisse haben.

Das bedingt aber, dass wir die Mathematik nicht in möglichst eleganter Form (wie es die Mathematiker lieben) darstellen können, sondern mit reichlich begleitendem Erläuterungstext. Wir werden die übliche Gliederungsform in Definition-Satz-Beweis beibehalten. Eilige Leser können den Zwischentext überspringen.

1.2 Ein spielerisches Experiment

Wir beginnen die Lehrveranstaltung mit einem kleinen Experiment: Der Vortragende präsentiert den Hörern ein Blatt mit einer Liste von Wörtern, die ohne erkennbare Regelmäßigkeit zusammengestellt sind. Er fotokopiert das Blatt und gibt Original und Kopie an Hörer mit der Aufforderung, einige der Wörter nach Belieben durch andere zu ersetzen. Die so manipulierten Kopien werden ihrerseits kopiert und mit diesen Kopien wird entsprechend verfahren, bis schließlich eine größere Zahl von Kopien, im Beipiel 15, im Umlauf sind. Die Daten sind in Abbildung 1.1 abgedruckt.

Ist es nun möglich, aus den Listen die Entstehungsgeschichte und vielleicht sogar die Ausgangsliste zu rekonstruieren? Hauptanhaltspunkt für eine solche Rekonstruktion werden sicher diejenigen Wörter sein, die auf mehreren Listen vorkommen. Aber wie ist vorzugehen, welche Argumente haben Beweiskraft, welche sind plausibel?

Wir werden diese Fragen aufgreifen, zuvor aber eine viel einfachere Frage stellen: Wie wird ein Ergebnis der Rekonstruktion aussehen? Vom Ablauf des Experimentes haben wir eine einfache Vorstellung, nämlich die eines **Stammbaumes**, wie er exemplarisch in Abbildung 1.2 dargestellt ist. Gelesen wird ein solches Diagramm von links nach rechts[1]. Die linke Ecke ist die **Wurzel** des Baumes und symbolisiert den Ausgangszustand, in unserem Falle also die Urliste. Die Verzweigungspunkte stellen die Zwischenzustände dar, die Endpunkte ganz rechts werden **Blätter** genannt und repräsentieren die Ergebnisse, also die Listen, die zum Schluss vorhanden sind. Die Kanten symbolisieren die Änderungen. In einem so einfachen Diagramm ist die Länge der Kanten übrigens ohne Bedeutung.

Die geforderte Rekonstruktion wäre erfolgreich, wenn es uns gelänge, die verschiedenen Listen derart an die Ecken des Baumes einzutragen, wie es dem tatsächlichen Ablauf des Experimentes entspricht. An den Kanten könnten dann die jewei-

[1] Diese etwas gewöhnungsbedürftige Zeichenrichtung erleichtert das Beschriften.

ebenso	scharf	kammer
backen	pracht	nennen
brille	daraus	mutter
pastor	blumen	schwan
anfang	ramsch	nachts
ungern	kirsch	kuchen
wetter	lodern	keller

ebenso	intern	sommer
dusche	pracht	nennen
läufer	schnee	lieben
pastor	probat	schwan
anfang	ramsch	europa
ungern	zeiger	kuchen
phonem	montag	humbug

ebenso	scharf	kammer
backen	pracht	nennen
obacht	daraus	mutter
pastor	blumen	schwan
anfang	ramsch	nachts
ungern	rennen	kuchen
wetter	lodern	weiche

ebenso	salzig	kabuff
dusche	raster	nennen
obacht	daraus	mutter
pastor	probat	ameise
muster	ramsch	laufen
spiele	potenz	kuchen
matsch	lodern	humbug

ebenso	scharf	kabuff
dusche	syntax	nennen
obacht	sexist	mörder
pastor	probat	schwan
anfang	ramsch	futter
fehler	bilder	kuchen
matsch	lodern	humbug

ebenso	scharf	kabuff
garten	raster	nennen
obacht	daraus	vater
pastor	probat	schwan
hunger	ramsch	laufen
spiele	herbst	kuchen
neubau	lodern	humbug

ebenso	intern	herbst
spülen	pracht	heulen
läufer	daraus	mutter
blumen	probat	schwan
anfang	ramsch	zufall
ungern	lachen	kuchen
phonem	lodern	humbug

syntax	scharf	sommer
backen	pracht	nennen
obacht	daraus	mutter
wasser	proton	praxis
anfang	ramsch	nachts
medium	rennen	kuchen
phonem	lodern	welche

ebenso	scharf	kabuff
kiefer	raster	nennen
obacht	daraus	mutter
pastor	geburt	schwan
muster	ramsch	laufen
spiele	herbst	kuchen
neubau	lodern	gleich

ebenso	scharf	kabuff
dusche	syntax	nennen
obacht	sonnen	mutter
pastor	probat	schwan
anfang	ramsch	auster
spiele	bilder	kuchen
matsch	lodern	humbug

ebenso	intern	winter
dusche	pracht	nennen
läufer	daraus	mutter
blumen	probat	schwan
anfang	ramsch	zufall
ungern	falter	roller
phonem	lodern	humbug

ebenso	scharf	sommer
backen	achten	nennen
obacht	daraus	mutter
halten	blumen	praxis
anfang	ramsch	nachts
ungern	rennen	kuchen
phonem	lodern	welche

ebenso	scharf	kabuff
garten	raster	nennen
obacht	daraus	mutter
pastor	probat	schwan
muster	ramsch	laufen
spiele	herbst	kuchen
neubau	lodern	humbug

ebenso	scharf	kammer
backen	pracht	nennen
obacht	daraus	mutter
pastor	blumen	schwan
anfang	ramsch	nachts
ungern	rennen	kuchen
wetter	lodern	keller

ebenso	scharf	sommer
backen	pracht	nennen
obacht	daraus	mutter
wasser	blumen	praxis
anfang	ramsch	nachts
ungern	rennen	kuchen
phonem	lodern	welche

Abbildung 1.1: Die Daten des Fotokopierexperiments.

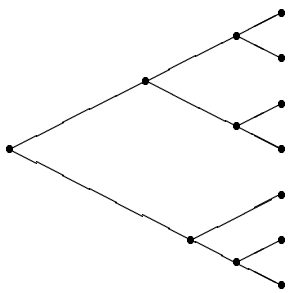


Abbildung 1.2: Ein (unbeschriftetes) Stammbaumdiagramm.

ligen Änderungen vermerkt werden. Möglicherweise geben wir uns aber auch mit weniger zufrieden, nämlich damit, bloß die Ergebnislisten den Blättern eines Baumes zuzuordnen, von dem wir argumentieren können, dass er die Entwicklung richtig wiedergibt. Ein solcher Nachweis ist oft leichter als die vollständige Rekonstruktion der Zwischenzustände.

Es gibt keinen besonderen Grund, warum gerade das abgebildete Diagramm den Verlauf unseres Experimentes richtig widerspiegeln sollte. Wir haben es als ein Beipiel von vielen möglichen angegeben. Tatsächlich wächst die Anzahl der Wurzelbäume sehr rasch mit der Eckenzahl. Schon für wenige Blätter ist sie so groß, dass sich ein Durchprobieren aller Möglichkeiten verbietet.

Natürlich ist unser Experiment nicht der eigentliche Anlass, das Stammbaummodell unter die Lupe zu nehmen. Uns interessiert die Frage, ob das Modell auch zur Beschreibung anderer, weit wichtigerer Abstammungssachverhalte taugt. Unser Hauptinteresse gilt dabei der Abstammungsgeschichte der **Sprachen**. Das Stammbaummodell ist in der vergleichenden Sprachwissenschaft alles andere als neu; es ist über ein Jahrhundert ausgiebig diskutiert worden. Man kann auch nicht sagen, dass es derzeit besonders aktuell ist. Viele Sprachwissenschaftler halten die Beschäftigung damit für wenig dringlich, viele bezweifeln, dass das Baummodell überhaupt zur Beschreibung von Sprachgeschichte taugt.

Ganz anders ist die Situation in der Biologie. Dort hat die **phylogenetische Rekonstruktion**[2] Konjunktur. Darunter versteht man die Bemühung, die Entwicklungsgeschichte der Lebewesen aus den heute bestehenden Arten zu erschließen. Ungeheuren Schwung hat dieses Vorgehen durch die modernen Techniken erfahren, den genetischen Code, also die Erbinformation der Lebewesen, zu entschlüsseln. Der Vergleich zu unserem Wortlistenexperiment ist leicht gezogen: Wenn es uns gelingt, mit Hilfe von ein paar Dutzend Wörtern etwa ebensoviele Listen in einen Abstammungsbaum zu ordnen, sollte es dann nicht wenigstens prinzipiell möglich sein, anhand von DNA-Ketten, die jeweils Information in der Gigabyte-Größenordnung haben, die gesamte Evolution von einigen hunderttausend Arten auszurechnen? Welch ein kühnes Projekt!

In der Biologie spielt das Stammbaummodell eine ganz dominante Rolle[3]. Das ist zumindest bei der ungeschlechtlichen Vermehrung der Einzeller sehr naheliegend. Wieder können wir die Analogie zum Kopierexperiment leicht herstellen: Bei der Vermehrung durch Zellteilung wird in der Tat eine Kopie des jeweiligen Erbgutes erstellt und an die Tochterzelle weitergegeben.

Allerdings fällt auch ein bemerkenswerter Unterschied zum Kopierexperiment auf: Unsere Fotokopien hatten wir nach jedem Kopiervorgang zum Ändern freige-

[2] phylon = der Stamm, genesis = die Entstehung.

[3] Auf Einschränkungen, Modifikationen und Verfeinerungen kommen wir noch zu sprechen.

geben; Veränderungen im Erbgut (Mutationen) hingegen sind selten im Vergleich zur Anzahl der Zellteilungen. Auch beim Kopierexperiment ist es nicht völlig ausgeschlossen, dass physisch verschiedene Wortlisten identischen Inhalts kursieren, das ist aber sicher die Ausnahme. Bei der Zellteilung erwarten wir im Kontrast dazu viele Individuen mit praktisch identischem Erbgut. Das ist ja auch hochwillkommen, denn Lebewesen, insbesondere Einzeller, neigen dazu, in großer Zahl aufzutreten, so dass ein Stammbaum der Individuen doch zuviel der Ehre wäre. Man wird sich also bemühen, Lebewesen mit identischem Erbgut oder auch nur charakteristischen Gemeinsamkeiten zusammenzufassen, zu Arten, Gruppen, Sorten. Dann ist es aber nicht ohne weiteres einsichtig, dass sich die Stammbaumvorstellung auch auf die Abstammung solcher Arten voneinander sinnvoll übertragen lässt.

Probleme mit dem Stammbaummodell auf der Ebene der Individuen selbst ergeben sich bei geschlechtlicher Fortpflanzung. Es ist nun einmal so, dass jeder Mensch zwei Eltern hat, und in seinen Erbmerkmalen Anteile der Mutter und des Vaters vereinigt. Man könnte das zum Anlass nehmen, das Stammbaummodell als für die Beschreibung der menschlichen Abstammung ungeeignet anzusehen, was natürlich absurd wäre, denn gerade da stammt die Idee ja her. Die Vorstellung vom „arbor consanguinitatis“ ist offenbar schon alt, und das Wort von der „Wurzel Jesse“ zeigt, dass diese Vorstellung vom Familienstammbaum durchaus auch in zeitlicher Vorwärtsrichtung gemeint ist. Den Genealogen war dabei keineswegs entgangen, dass der Mensch sich nicht durch Zellteilung vermehrt. Sie haben die Stammbaumdenkweise einfach dadurch gesichert, dass sie sich auf die männlichen Nachfahren beschränkten. Es ist in der Tat sehr plausibel, dass sich die männliche Abstammungslinie durch einen Stammbaum treffend beschreiben lässt.

Ebenso offensichtlich ist es, dass dieser Stammbaum nicht jede Familienähnlichkeit gut zu erklären vermag. Natürlich können zwei Menschen eng miteinander verwandt und aus diesem Grunde einander sehr ähnlich sein, ohne einen nahen gemeinsamen männlichen Vorfahren zu haben. Familienähnlichkeit taugt also nur sehr bedingt zur Beschreibung oder gar Rekonstruktion der Abstammung. Die Genealogen stört das wenig, weil sie andere Möglichkeiten benutzen, Stammbäume zu ermitteln.

Kurzum: Merkmalgleichheit bzw. Ähnlichkeit sowie Nähe in Stammbaum sind nicht identisch. Sie haben ohne Zweifel etwas miteinander zu tun, und wir werden dem nachgehen, aber man darf sie nicht miteinander verwechseln. Hätten wir beim Kopierexperiment den Teilnehmern Zeit gegeben, die Listen nach Herzenslust zu verändern, so hätte sich schon nach wenigen „Generationen“ der Wortvorrat völlig gewandelt und die Listen hätten schließlich womöglich gar keine gemeinsamen Wörter mehr enthalten. Das würde eine Rekonstruktion aussichtslos erscheinen lassen, wäre aber kein Grund, die Baumstruktur der Abstammungsgeschichte zu bezweifeln.

Wir werden diese beiden Themenbereiche deshalb trennen. Zuerst untersuchen wir die Frage, welche Annahmen ein Stammbaummodell rechtfertigen und welche es in Frage stellen. Dann kümmern wir uns darum, auf welche Weise aus Merkmalen auf oder gegen Stammbäume geschlossen werden kann.

Es ist also nicht unser Ziel, die Frage nach dem „richtigen“ Stammbaum für die Sprachen zu entscheiden. Uns geht es darum, *ob* diese Modellvorstellung trägt und wie weit. Und wenn sich herausstellen sollte, dass sie nicht geeignet ist, die Abstammung der Sprachen zu beschreiben? War dann unsere Bemühung umsonst?

Wir glauben, dass das nicht so ist. Die eigentlich spannende Frage ist unserer Meinung nach nämlich die nach geeigneten Alternativen zum Stammbaummodell. Wir glauben, dass solche Alternativen nur dann gefunden werden können, wenn die Vor- und Nachteile des Baummodells sorgfältig untersucht sind.

Und was ist nun die richtige Lösung zum Fotocopierexperiment? Das wird vorläufig nicht verraten. Die Daten liegen vor, und jeder darf sich daran versuchen.

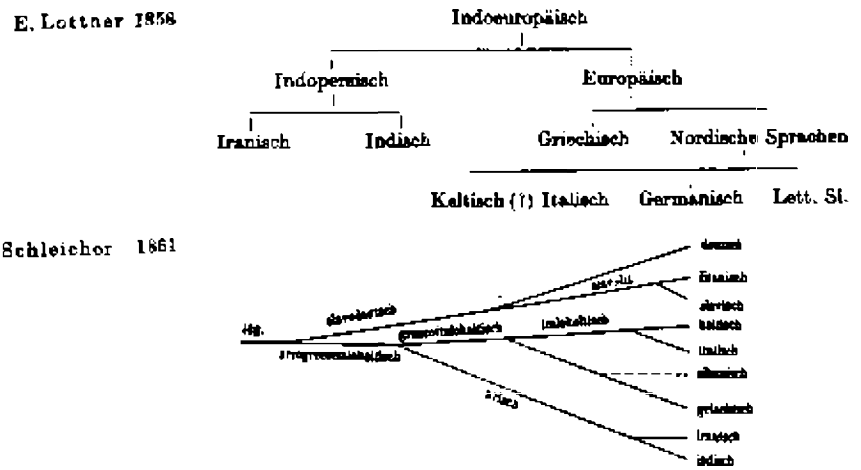


Abbildung 1.3: Stammbäume um 1860.

1.3 Sprachvergleich und Stammbäume

Die Aufgabe des Sprachvergleichs ist nicht unsere willkürliche Idee, sondern Forschungsgegenstand eines kleinen, aber traditionsreichen Wissenschaftszweiges, der *Vergleichenden Sprachwisenschaft*. Forschergruppen an gut zwei Dutzend Universitäten beschäftigen sich mit diesem Thema, meist im deutschsprachigen Raum, also in Deutschland, Österreich, der Schweiz, aber auch in Italien und den USA. Traditionell ist „Vergleichende Sprachwissenschaft“ synonym zu *Indogermanistik*, denn das vergleichende Studium der Sprachen und ihrer Geschichte hat sich hauptsächlich auf die indogermanischen Sprachen konzentriert. Über die Hälfte der heute gesprochenen Sprachen gehört dieser Sprachfamilie an. Zur Entstehung der Vergleichenden Sprachwissenschaft lese man Abschnitt 4.1.

Die Vorstellung, dass sich die Entstehungsgeschichte der Sprachen durch einen Stammbaum darstellen lässt, ist Mitte des neunzehnten Jahrhunderts in den Vordergrund getreten. Berühmt ist der von August Schleicher 1853 publizierte Stammbaum, weitere Beisiele aus dieser Zeit zeigt Abbildung 1.3. Die Sprachwissenschaften folgten damals nicht dem Vorbild der Biologie, denn die biologische Evolutionstheorie wurde in dieser Zeit ebenfalls gerade erst formuliert. Man bedenke, dass Darwins Werk *On the Origin of Species by Means of Natural Selection* erst 1859 erschien und lange Zeit tief umstritten war. Der Gedanke an die Fortentwicklung der Sprachen war ja auch viel näherliegend und leichter zu akzeptieren als der von der Veränderlichkeit und Unvollkommenheit der Schöpfung. Möglicherweise war der Einfluss sogar andersherum: Das Stammbaummodell in seiner klassischen naturwissenschaftlichen Ausprägung findet sich vor allem bei Ernst Haeckel, der dazu schreibt: „Das ganze natürliche System der Pflanzen erscheint von diesem Gesichtspunkte [d.h., dem der Deszendenzlehre] aus als ein grosser Stammbaum, und lässt sich eine jede genealogische Tabelle, am anschaulichsten unter dem Bilde eines weit verzweigten Baumes darstellen, dessen ganz einfache Wurzel in der fernsten Vergangenheit verborgen liegt.“

Haeckel beruft sich dabei mehrfach auf Schleicher und schreibt beispielsweise: „Dieser interessante Parallelismus in der divergenten Entwicklung der Sprachformen und der Organismen-Formen ist in sehr einleuchtender Weise von einem unserer ersten vergleichenden Sprachforscher erörtert worden, von dem genialen August Schleicher, der namentlich den Stammbaum der indogermanischen Sprachen in der scharfsinnigsten Weise phylogenetisch entwickelt hat.“ Es ist also ersichtlich, dass

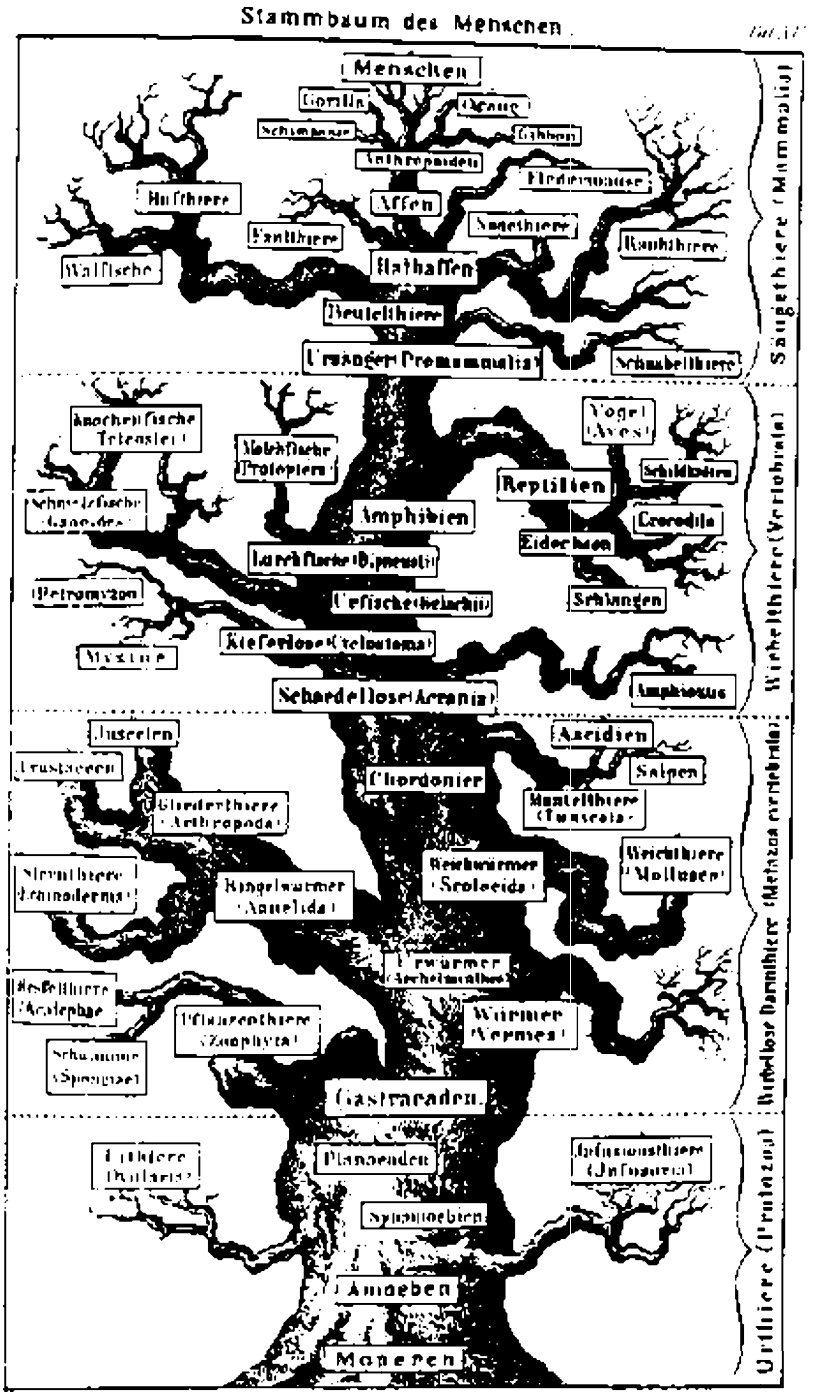


Abbildung 1.4: Stammbaum des Menschen nach E.Haeckel (1884)

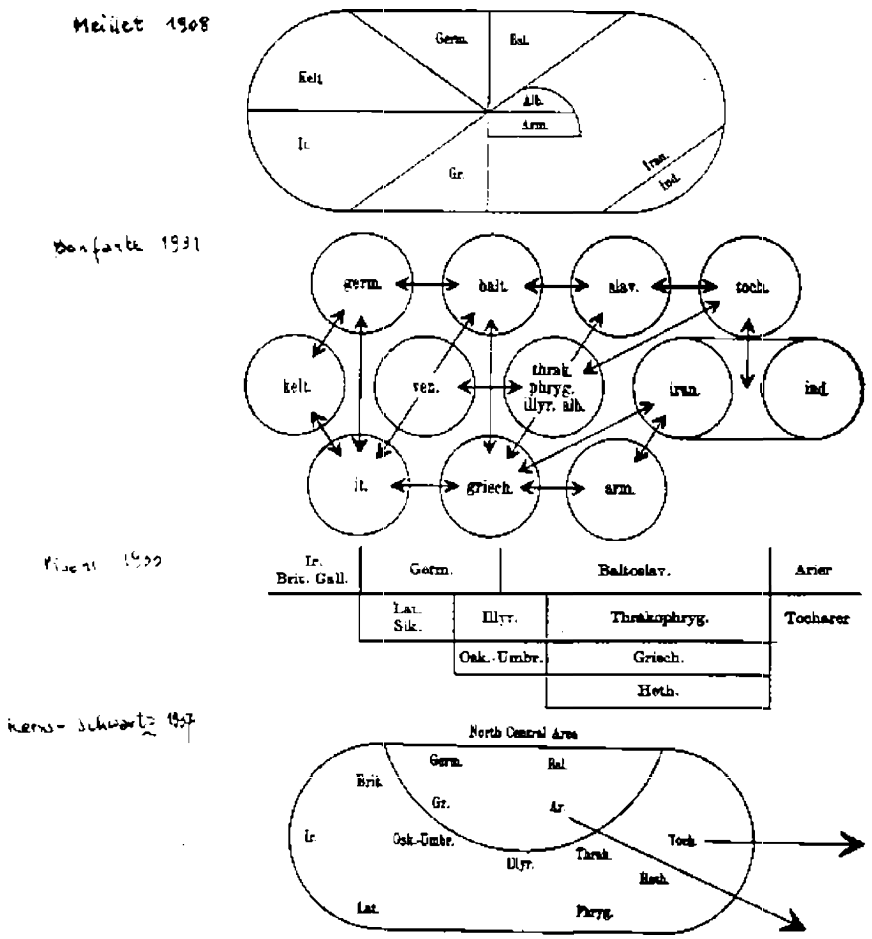


Abbildung 1.5: Andere Darstellungsformen von Sprachverwandtschaft.

es zu Schleichers Zeiten eine fruchtbare wechselseitige Beeinflussung von Sprachwissenschaft und Naturwissenschaften gegeben hat.

Die Sprachwissenschaft ist aber nicht bei dem Stammbaummodell allein geblieben, sondern hat eine ganze Reihe anderer Möglichkeiten ausgelotet, Verwandtschaften darzustellen. Abbildung 1.5 zeigt Beispiele. Diese Beispiele sind wohlüberlegt und ausgefeilt, aber sie stiften auch Verwirrung, besonders wenn man sie (wie wir es tun) ohne eine genaue Leseregel vorstellt. Was bedeuten diese Graphiken, welchen Wahrheitsgehalt haben sie? Stellen sie Metaphern, Veranschaulichungen und Merkregeln dar, oder enthalten sie konkrete sprachwissenschaftliche Aussagen? Kann man sagen, dass eine solche Graphik *richtig* oder *falsch* ist, und welche Argumente muss man vorbringen, um dies zu belegen? Beim Stammbaummodell stellten sich diese Fragen nicht sogleich. Jedenfalls wenn wir an eine so einfache Situation wie unser Fotokopierexperiment denken, so sind wir doch recht sicher, dass sich die Ereignisse, die zu unseren Daten in Abbildung 1.1 geführt haben, so zugetragen haben, dass sie sich in einen Stammbaum einordnen lassen.

Die Verwirrung wird noch dadurch gesteigert, dass eine ganz andere Verwendung von Baumstrukturen in den Sprachwissenschaften vorkommt, die man leicht mit der vom Stammbaum verwechseln kann, weil sie ihr äußerlich gleicht. Ein Beispiel zeigt Abbildung 1.6.

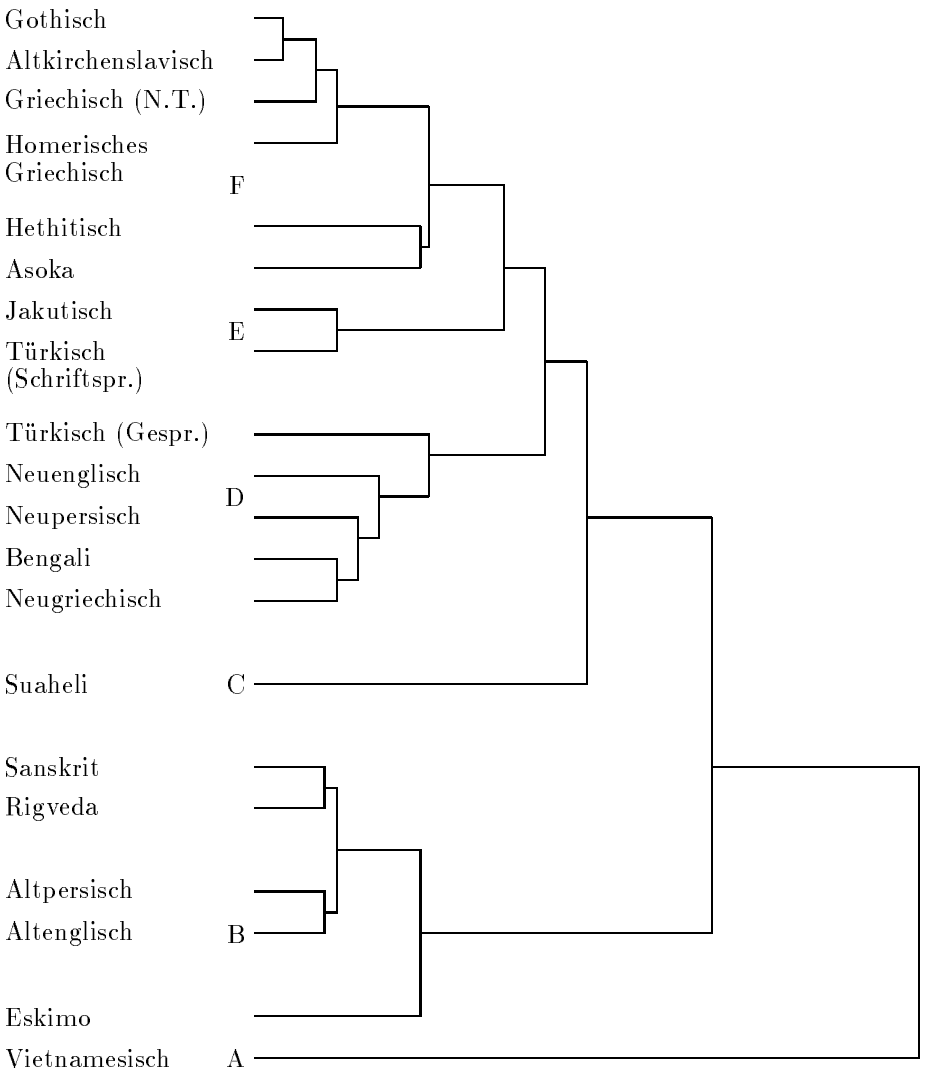


Abbildung 1.6: Klassifikation von 20 Sprachen nach Altmann und Lehfeldt[2] (Längenwiedergabe nicht exakt).

Abbildung 1.7: Stammbaum der indoeuropäischen Sprachen nach Cavallo-Sforza.

Abbildung 1.8: Stammbaum der indoeuropäischen Sprachen nach Ringe und Warnow.

1.4 Das Stammbaummodell heute

Bei aller Experimentierfreude und Kritik im Detail hat die Vergleichende Sprachwissenschaft das Stammbaummodell keineswegs verworfen und vergessen. In unterschiedlicher Ausprägung, aber mit großer Selbstverständlichkeit tritt es immer wieder hervor. Wir erwähnen zwei Beispiele: einen populärwissenschaftlichen Aufsatz von Cavallo-Sforza, dem wir Abbildung 1.7 entnommen haben, und „The Computational Historical Linguistics Project“ an der University of Pennsylvania, deren Homepage (http://linc.cis.upenn.edu/ histling/home.html) die in Abbildung 1.8 wiedergegebene Graphik schmückt.

Wir haben aber den Eindruck, dass die Vergleichende Sprachwissenschaft beim Stammbaummodell gegenüber der Biologie ins Hintertreffen geraten ist. Das hat verschiedene Gründe. Die Sprachwissenschaft hat schlechte Erfahrungen gemacht: davon wird später noch ausführlich berichtet. Umgekehrt gab es in der Biologie fördernde Faktoren, aber auch Auseinandersetzungen. Das führte dort dazu, dass sich eine strenge, ausgefeilte Methodik durchgesetzt hat, die der Systematiker Willi Hennig [10] formuliert hat. Mit der Entschlüsselung des genetischen Codes hatte die Biologie dann plötzlich außerordentlich trennscharfe Daten zur Verfügung, die in Kombination mit der fortgeschrittenen Methodologie eine Welle von Erkenntnissen ermöglicht haben.

Wir haben die Hoffnung, dass sich manches von dem Schwung, den die Phylogenetische Systematik der Biologie hat, auch für sprachwissenschaftliche Fragestellungen nutzen lässt. Allerdings sind die biologischen Methoden oft nicht ohne Modifikation anwendbar, denn sie sind zu spezifisch.

Kapitel 2

Eine Formalisierung des Stammbaummodells

Die folgenden Zeilen beschreiben die Annäherung eines Strukturmathematikers an das Stammbaummodell der historischen Sprachwissenschaft. Wir versuchen dabei, ganz genau herauszuarbeiten, welche Annahmen dem Modell vorausgehen. Dazu werden die Annahmen präzise formuliert, und zwar mit einer Pingeligkeit, die demjenigen, der diese „axiomatische" Methode des Herangehens nicht gewohnt ist, als übergenau und ermüdend erscheinen muß. Durch die Zuspitzung und Atomisierung werden nämlich die Voraussetzungen in Einzelschritte zerlegt, die derart abgemagert sind, daß sie wie Trivialitäten erscheinen, über die zu reden sich nicht lohnt.

Darin besteht aber gerade die Provokation. Die Darstellung ist nämlich so beschaffen, daß die „Trivialitäten" zusammengenommen einem Stammbaummodell logisch äquivalent sind. Wer also akzeptiert, daß über das Modell analytisch geredet werden kann und dabei die seit Aristoteles gängigen logischen Schlußregeln nicht verwirft, der hat nur die Wahl, dem Stammbaummodell zuzustimmen oder eine der herausgearbeiteten Modellannahmen als nicht haltbar anzugreifen.

Wir verfolgen diese Darstellung in der Absicht, das Modell

- *besser kritisierbar zu machen,*

  indem wir genau unterscheiden, was Annahme und was Folgerung ist und auch darlegen, welche Argumente *gegen* das Modell sprechen können, es

- *besser benutzbar zu machen,*

  indem wir deutlich machen, welche Folgerungen ohne zusätzliche Annahmen abgeleitet werden können, und es

- *flexibler zu machen,*

  indem wir zeigen, was bei Änderungen der Modellannahme anders wird und was nicht.

Die Modellierung geschieht zuerst *qualitativ*, was bedeuten soll, daß Zahlenwerte für Dauern, Ähnlichkeiten oder Wahrscheinlichkeiten nicht betrachtet werden, sondern die reine Ordnungsstruktur im Vordergrund steht. Eine spätere Anreicherung des Modells ist damit nicht ausgeschlossen.

2.1 Ein synoptisches Axiomensystem

Zuerst wählen wir einen Sprachrahmen. Dazu gehen wir von der Annahme aus, daß eine Aussage wie

„Das Französische stammt vom Lateinischen ab“

den meisten Sprachwissenschaftlern sinnvoll vorkommt. Wir meinen damit nicht, daß dieser Satz *wahr* ist. Das ist er sicherlich nicht im Sinne einer unmittelbaren Wahrnehmung, denn „das Französiche“ und „das Lateinische“ sind ebensowenig beobachtbare Gegenstände wie der Vorgang der Abstammung. Es handelt sich um Abstraktionen, Konstrukte. Der Grund, aus welchem wir Aussagen der Form

„Sprache A stammt von Sprache B ab“

als das wesentliche Ausdruckmittel für das folgende Modell wählen, ist ein anderer. Wir erwarten, daß sich Sprachwissenschaftler mittels solcher Aussagen über Stammbäume verständigen können. In diesem Sinne ist die Aussage „Das Lateinische stammt vom Japanischen ab“ schon deshalb innerhalb des Sprachrahmens, weil ihr die allermeisten Sprachwissenschaftler sogleich einen Wahrheitswert (nämlich „falsch“) zuschreiben.

Damit Wissenschaftler mit Hilfe derartiger Aussagen kommunizieren können, müssen sie diese Sätze keineswegs in genau der gleichen Weise verstehen. Das wäre eine kaum erfüllbare Forderung. Es genügt, wenn sie sich auf einige Regeln des gemeinsamen Gebrauchs einigen. Genau dies ist die Aufgabe der nun folgenden Axiome.

2.1.1 Die Annahmen zum qualitativen Baummodell

Die Beschreibung der Modellannahmen geschieht verbal, zusätzlich wird eine mathematische Formulierung angegeben (die meist sehr viel kürzer ist). Wir nennen dies ein **synoptisches Axiomensystem**. Jede der Annahmen wird erläutert; im folgenden sind nur die hervorgehobenen Sätze 1) bis 7) Modellannahmen, der Rest ist Erläuterung. Allerdings erhält oft die Kurzfassung erst durch die Erläuterung ihren präzisen Sinn. Auch die mathematische Fassung ist so zu verstehen: sie erläutert und präzisiert die verbale Version.

Die nun folgenden sieben Bedingungen sind von sehr unterschiedlicher Qualität. Während die ersten beiden den Sprachrahmen festlegen und dadurch ganz einschneidend sind, treffen 3), 4) und 7) eher unwichtige Vereinbarungen, die zur Glättung der Formulierung dienen. 5) und 6) enthalten die wichtigsten Strukturannahmen.

Jeweils am Ende eines Abschnittes geben wir hinter dem Zeichen „**M:** “ die mathematische Fassung der diskutierten Modellannahme an.

1. Das Stammbaummodell betrifft Sprachen.

Wir setzen also voraus, daß man sinnvoll von Sprachen reden kann und daß man die Sprachen, von denen die Rede ist, unterscheiden kann. Wir sagen dazu nicht, wie Sprachen definiert werden. Wir unterscheiden auch an dieser Stelle nicht zwischen realen und rekonstruierten oder zwischen lebenden und toten Sprachen. Auf die innere Struktur der Sprachen kommt es uns nämlich in dieser Phase der Modellbildung gar nicht an. Tatsächlich läßt sich das Stammbaummodell auch auf ganz andere Situationen anwenden, etwa auf biologische Arten oder auf Handschriftenfamilien.

Unser Modell betrifft nicht *„alle“* Sprachen, sondern eine willkürliche Auswahl (die nicht von uns, sondern vom jeweiligen Anwender getroffen wird). Diese Auswahl ist nicht als endgültig zu verstehen, sondern kann nach Belieben geändert werden. Wenn im folgenden von „Sprachen“ die Rede ist, sind stets Sprachen der jeweilig betrachteten Auswahl gemeint.

Die mathematische Methode strebt an, daß nur diejenigen Annahmen benutzt werden, die auch formuliert sind. Es sollte an dieser Stelle deutlich sein, was hier

nicht vorausgesetzt wird: daß nämlich Sprachen all die Eigenschaften haben, die vorauszusetzen wir gewohnt sind. Wir verlangen hier *nicht*, daß Sprachen eine einheitliche Struktur haben, daß Menschen nur eine Sprache haben usw. usw. Es kann sein, daß solche Bedingungen durch spätere Modellannahmen noch gefordert werden. Aus der oben formulierten Annahme folgt all dies jedenfalls nicht.
M: Die betrachteten Sprachen bilden eine Menge (Zeichen: S).

2. Sprachen können voneinander abstammen.

Für je zwei Sprachen soll sinnvollerweise gefragt werden können, ob sie voneinander abstammen. Genauer: zu je zwei Sprachen soll die Aussage „die erste Sprache stammt von der zweiten ab“ sinnvoll, das heißt richtig oder falsch sein. Es muß nicht *bekannt* sein, ob die eine Sprache von der anderen abstammt, es muß lediglich prinzipiell entweder zutreffen oder nicht. Statt „A stammt von B ab“ sagen wir auch „B ist ein Vorfahr von A“
M: „Stammt ab von“ ist eine Relation auf S. Zeichen: „$<$“.

3. Wir betrachten nur endlich viele Sprachen.

Diese Vereinbarung ist keineswegs so selbstverständlich und ihr Gegenteil keineswegs so absurd, wie man vorschnell meinen könnte. Denn die realen Sprachen leben ja in der Zeit und ändern sich unablässig. Würde man eine Modellierung anstreben, die alle Zeitpunkte unterscheidet, so hätte man wohl unendlich viele Sprachen zu berücksichtigen.

Von einer solchen Modellierung raten wir aber ab. Nicht weil sie wesentlich schwieriger wäre (sie ist es wohl nicht), sondern weil sie Probleme suggeriert, wo keine sind. Man beachte, daß wir damit nicht die eher esoterische Frage zu beantworten versuchen, ob es denn nun unendlich viele Sprachen „gibt“, sondern lediglich festlegen, daß wir eine endliche Auswahl betrachten wollen. Das kann keine spürbare Einschränkung bedeuten: Selbst wenn wir die Zeit in Sekundenabschnitte unterteilen würden, bekämen wir endlich viele Zeitpunkte seit dem Beginn des Sprachvermögens, und damit endlich viele Sprachen (oder die sicher problematische Annahme, daß es unendlich viele Sprachen zu einem festen Zeitpunkt gegehen haben muß). Man beachte im übrigen, daß in der rein qualitativen Grundfassung des Stammbaummodells von „Zeit“ nicht die Rede ist.
M: S ist endlich.

4. Keine Sprache stammt von sich selbst ab.

Das ist nur eine Klarstellung über die Verwendung der Relation „stammt ab“, und ohne wirkliche Bedeutung für die Modellbildung. Die folgenden Regeln formulieren sich mit dieser leichter.
M: „Stammt ab von“ ist irreflexiv.

5. Abstammung vererbt sich.

Genauer: Wenn eine Sprache von einer zweiten abstammt und diese von einer dritten, dann stammt die erste auch von der dritten ab.
M: „Stammt ab von“ ist transitiv. $<$ ist also eine strikte Ordnung auf S.

6. Abstammung ist rückwärts unverzweigt.

Genauer: Stammt eine Sprache A sowohl von der Sprache B als auch von der Sprache C ab (B und C verschieden), dann stammt B von C oder C von B ab.

Diese Bedingung ist entscheidend für die Baumstruktur, weil sie Hybridisierungen verbietet. Modifikationen des Stammbaummodells sollten unserer Meinung nach am ehesten hier ansetzen. Wir warnen aber davor, diese Annahme vorschnell zu verwerfen. Wer schon allein deshalb, weil eine Sprache Einflüsse aus verschiedenen Quellen ausgesetzt ist, diese Annahme für unzutreffend hält, hat vermutlich den einleitenden Abschnitt nicht gelesen und trennt Modell und Realität nicht sorgfältig genug.
M: $a < b$, $a < c$ und $b \neq c$ impliziert $b < c$ oder $c < b$.

Die Annahmen 1 – 6 implizieren bereits, daß die Sprachen bezüglich der Abstammung waldartig geordnet sind, also durch (möglicherweise mehrere) Stammbäume. Wir haben nichts gefordert, was gewährleisten könnte, daß diese Bäume miteinander zusammenhängen und erhalten deshalb einen „Stammwald". Der Einfachheit halber betrachten wir die Bäume dieses Waldes einzeln. Mit anderen Worten: wir betrachten jeweils nur eine Auswahl von Sprachen, die allesamt miteinander verwandt sind.

7. Alle betrachteten Sprachen sind miteinander verwandt.

Dabei ist folgendermaßen erklärt, wann zwei Sprachen miteinander verwandt sind: Sprachen, die voneinander abstammen, sind miteinander verwandt, und wenn zwei Sprachen beide mit einer dritten verwandt sind, dann sind sie auch untereinander verwandt.

Zugegebenermaßen handelt es sich bei der letztgenannten Annahme etwas um Augenwischerei. Man macht sich nämlich rasch klar, daß aufgrund der anderen Annahmen gilt: Zwei Sprachen sind genau dann im beschriebenen Sinne verwandt, wenn sie einen gemeinsamen Vorfahren haben. Annahme 7 ist deshalb gleichbedeutend zur Annahme *Es gibt eine Sprache, von der die anderen abstammen.* Wir mochten diesen Satz aber nicht zum Axiom erheben, weil der schnelle Leser, der nur das Fettgedruckte liest, dann vorschnell annimmt, daß hier ein gemeinsamer Ursprung aller Sprachen der Welt postuliert wird. Es folgt bereits aus 6), daß die betrachteten Sprachenmenge in Stammbäume zerfällt; das Axiom stellt nur klar, daß wir uns auf einen einzelnen Baum konzentrieren. Die Stammsprache der Sprachen dieser Auswahl wird im folgenden gelegentlich als *Wurzelsprache*[1] bezeichnet.
M: Es gibt ein größtes Element der Ordnung $(S, \leq)$.

Zur besseren Referenz stellen wir die beiden Axiomensysteme in Abbildung 2.1 noch einmal in Kurzfassung zusammen. Eine komprimierte Fassung der mathematischen Version ist:

Definition 1 Unter einem **Stammbaum** verstehen wir eine endliche geordnete Menge $(S, \leq)$ mit größtem Element, welche die Bedingung 6 erfüllt. ◊

2.2 Stimmt das Modell?

Die oben vorgestellten Annahmen über Sprachen und ihre Abstammung beschreiben die Wirklichkeit nicht im Sinne unmittelbarer Anschauung. Sie sind idealisierend und begründen zusammen ein *Modell*. Man kann ein solches Modell auf zwei Weisen beeinflussen: Erstens indem man es formuliert bzw. ändert. Dies geschieht in der *Modellbildungsphase*, gewöhnlich streng getrennt von der zweiten Möglichkeit, die darin besteht, das Modell zu benutzen.

[1] In der Indogermanistik ist es üblich, Elemente der rekonstruierten indogermanischen Ursprache als *Wurzeln* zu bezeichnen. Diese Bedeutung des Wortes ist hier nicht gemeint.

Die verbale Fassung		Die mathematische Version
Das Stammbaummodell betrifft Sprachen.	1	Die betrachteten Sprachen bilden eine Menge (Zeichen: S).
Sprachen können voneinander abstammen.	2	„Stammt ab von“ ist eine Relation auf S. Zeichen: „$<$“.
Wir betrachten nur endlich viele Sprachen.	3	S ist endlich.
Keine Sprache stammt von sich selbst ab.	4	$<$ ist irreflexiv.
Abstammung vererbt sich.	5	$<$ ist transitiv.
Abstammung ist rückwärts unverzweigt.	6	$a < b$, $a < c$ und $b \neq c$ impliziert $b < c$ oder $c < b$.
Alle betrachteten Sprachen sind miteinander verwandt.	7	Es gibt ein größtes Element der Ordnung $(S, \leq)$.

Abbildung 2.1: Synoptische Axiome zum Stammbaummodell.

Zweitens, indem das Modell durch die Angabe von Daten konkretisiert wird, und zwar von solchen Daten, auf die die Modellannahmen angewendet werden können. Aufgabe des Modells ist es, die Konsequenzen sichtbar zu machen, die die Daten zusammen mit den Modellannahmen nach sich ziehen. Daten bestehen in unseren Fall aus Urteilen wie etwa „a ist eine Sprache“ oder „b stammt von c ab“. Daten kommen durch *Beobachtung* zustande, die sehr unterschiedlicher Natur sein können. Sie unterliegen, ebenso wie das Modell selbst, der *Bewertung*. Diese muß insbesondere dann greifen, wenn sich Widersprüche ergeben. Angestrebt wird, daß ein Modell mit den Daten kompatibel ist. Ergeben sich Inkonsistenzen, so ist dies Anlaß zur Revision. Gewöhnlich werden zuerst die Daten überprüft. Erweisen sich die Daten aber als korrekt und dennoch inkompatibel mit dem Modell, dann muß dieses verworfen oder angepaßt werden; die Anpassung kann darin bestehen, daß der Datenbereich eingeschränkt wird.

Ein akzeptiertes Modell kann auch als Prognoseinstrument benutzt werden. Führt nämlich eine hypothetische Datenmenge A zum Widerspruch mit dem Modell, so kann das als die Modellprognose gedeutet werden, daß die Datenmenge A nicht beobachtet wird. Eine falsche Prognose führt dann zum Verwerfen des Modells.

Der Wert eines Modells entscheidet sich an seiner prognostischen Kraft. Ein Modell, welches mit beliebigen Daten kompatibel ist, also immerzu „ja, ja“ sagt, ist nutzlos.

Ein Modell „stimmt“ also nicht; nur in den einfachsten Fällen wird man Modelle formulieren können, die als endgültig anzusehen sind. Gewöhnlich wird jedes Modell irgendwann an Grenzen stoßen, die eine Revision erforderlich machen, denn Modelle sind ja nicht eins mit der Wirklichkeit. Eine solche Revision bedeutet aber nicht unbedingt, daß das vorhergehende Modell wertlos geworden ist. Oft ist es für die Entwicklung und das Verständnis des besseren Modells unabdingbar, das einfachere Modell verstanden zu haben.

In diesem Sinne verstehen wir auch das hier beschriebene Stammbaummodell. Wir erwarten nicht, daß es die linguistischen Sachverhalte der Sprachenentwicklung genau beschreibt. Wir kennen aber auch kein ausgearbeitetes formales Modell, welches das Stammbaummodell überträfe. Unser Ziel ist es, durch exakte Modellformulierung und durch darauf gegründete Datenanalyse die Inkonsistenzen des Modells so genau herauszupräparieren, daß die Anforderungen an ein weitergehendes Modell sichtbar werden.

2.3 Diagramme und erste Folgerungen

Für jede Sprache ist die Menge der „Vorfahren“, also die Menge derjenigen Sprachen (unserer Auswahl), von denen die Sprache abstammt, *linear geordnet*, d.h. es gibt unter ihnen eine erste, von der alle anderen abstammen, eine zweite, von der alle bis auf die erste abstammen, und so fort. Aus dem Umstand, daß die von uns getroffene Auswahl nur endlich viele Sprachen enthält, folgt, daß jede Sprache außer der Wurzelsprache genau einen unmittelbaren Vorfahren hat (wir nennen dies die *Vatersprache*) und daß jede Sprache, die überhaupt Nachfahren in unserer Auswahl hat, auch Tochtersprachen hat, womit Sprachen gemeint sind, zu denen die erstgenannte Sprache die Vatersprache ist. Dies erlaubt es, einfache graphische Darstellungen und Sprachelemente einzuführen, die sich ohne zusätzliche Annahmen erklären lassen und die die Verständigung erleichtern.

Die oben modellierten Strukturen lassen sich gut durch die schon diskutierten *Baumdiagramme* wiedergeben. Dabei werden die Sprachen auf dem Papier durch kleine Kreise dargestellt, wobei die bezüglich der Abstammung „früheren“ Sprachen weiter links und die Nachfahren weiter rechts gezeichnet werden (der Baum hat die „Wurzel“ links). Jeder Kreis wird mit den Kreisen der Tochtersprachen durch nach rechts weisende Linien verbunden. Das bewirkt, daß man genau dann von einem Kreis längs eines linksweisenden Linienzuges zu einem zweiten gelangen kann, wenn die zugehörige Sprache von der zweiten abstammt. Linien, die Generationen überspringen, werden nicht eingezeichnet. Nach Bedingung 5 ergibt sich die von ihnen darzustellende Information mittels der anderen Linien. Das Weglassen macht das Bild besser lesbar. Im Diagramm sind die Stammbaum-Informationen vollständig ablesbar; ein Stammbaum kann also durch ein Diagramm angegeben werden, in dem die Kreise mit den Namen der Sprachen beschriftet sind.

Weil wir die „stammt ab von“-Relation mathematisch als strikte Ordnung modellieren, benutzen wir auch einfach das übliche Kürzel “$<$“ dafür. Entsprechend steht „$\leq$“ für „ist Nachfahr von oder gleich“, “$a > b$“ wird „a ist Vorfahr von b“ gelesen, und das Zeichen „$\geq$“ steht für „ist Vorfahr von oder gleich“. Die Vorfahren einer Sprache sind linear geordnet. Deshalb gibt es zu je zwei Sprachen einen *spätesten gemeinsamen Vorfahren*, ähnlich dem kleinsten gemeinsamen Vielfachen bei Zahlen. Wir verwenden dafür das Zeichen $\vee$, und definieren präziser:

> Für Sprachen a und b bezeichne $a \vee b$ diejenige Sprache v der Auswahl, die durch die folgenden Eigenschaften charakterisiert ist:
>
> 1. $a \leq v$ und $b \leq v$,
> 2. Es gibt keinen Nachfahren w von v mit $a \leq w$ und $b \leq w$.

Man beachte, daß die Existenz einer solchen Sprache nicht zusätzlich gefordert wird, sondern Konsequenz unserer Modellannahmen ist.

Es ergibt sich unmittelbar aus der Definition, daß $a \vee b = b \vee a$ und $a \vee a = a$ gilt. Außerdem überlegt man, daß es beim Iterieren auch auf Klammerungen nicht ankommt und $a \vee (b \vee c) = (a \vee b) \vee c$ für je drei Sprachen a, b, c richtig ist. Deshalb überträgt sich unsere Abkürzung auf mehr als zwei Sprachen: Ist $A :=$

$\{a_1, a_2, \ldots, a_n\}$ eine beliebige (nichtleere) Teilauswahl unserer Sprachenmenge S, dann ist $a_1 \vee a_2 \vee \ldots \vee a_n$ der späteste gemeinsame Vorfahr aller Sprachen aus A. Wir benutzen dafür auch die Kurzschreibweise $\bigvee A$. Speziell ist $\bigvee S$ die Wurzelsprache.

Hilfssatz 1 *Ist $A \subseteq S$ und $a \in A$ beliebig, dann gibt es ein $b \in A$ mit $\bigvee A = a \vee b$.*

Beweis Die Elemente der Form $a \vee r$, $r \in A$, sind sämtlich $\geq a$ und deshalb linear geordnet. Sei b ein Element, für das $a \vee b$ maximal unter solchen Elementen ist, dann gilt $a \vee b \geq a \vee r \geq r$ für alle $r \in A$, also $a \vee b \geq \bigvee A$ und damit $a \vee b = \bigvee A$. □

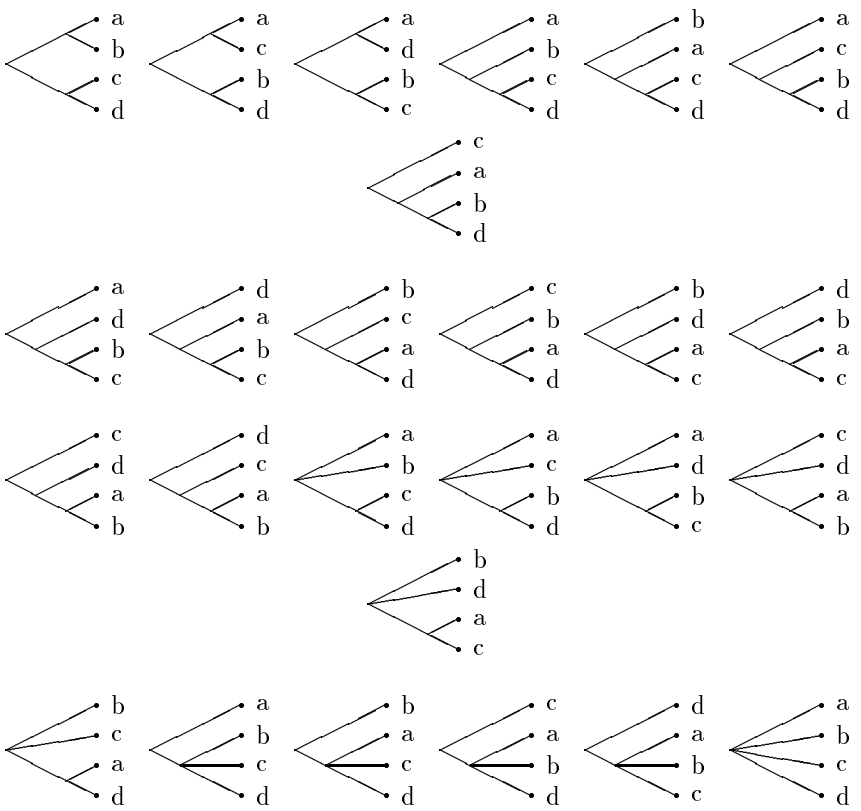


Abbildung 2.2: Die Anzahl möglicher Bäume ist schon für wenige Sprachen recht groß. Läßt man nur Aufspaltungen in jeweils zwei Tochtersprachen zu, so gibt es eine einfache Formel: die Anzahl der binären Wurzelbäume mit $n+1$ Blättern ist $1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2n-1)$, was annähernd gleich $(\frac{2n}{e})^n$ ist.

2.4 Bäume ohne Wurzel

Was wir in Definition 1 als *Stammbaum* eingeführt haben, würde man in der Mathematik einen **Wurzelbaum** nennen (engl. **rooted tree**), denn das ist der gängige Terminus des dafür zuständigen Arbeitsgebietes, der *Graphentheorie*. Weil deren Grundbegriffe so einfach, anschaulich und nützlich sind, erwähnen wir sie hier kurz.

In der Sprache der Graphentheorie besteht ein (schlingenloser, ungerichteter) **Graph** aus einer Menge V, deren Elemente man **Ecken** nennt, und einer Menge

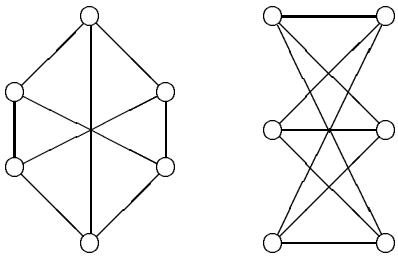


Abbildung 2.3: Beispiele von Graphendiagrammen

Abbildung 2.4: Beispiel eines Baumdiagramms

E, bestehend aus zweielementigen Teilmengen von V; diese werden die **Kanten** des Graphen genannt.

Man kann Graphen wie Stammbäume durch Diagramme veranschaulichen, bei denen die Ecken durch kleine Kreise auf dem Papier dargestellt sind und jede Kante durch eine Linie, die zwei solche Kreise verbindet, repräsentiert wird. Abbildung 2.3 zweigt zwei derartige Diagramme mit jeweils sechs Ecken und neun Kanten. Auf die genaue lage der Kreise kommt es dabei nicht an, sondern nur darauf, welche miteinander verbunden sind. Genauere Inspektion zeigt, dass diese beiden Diagramme den gleichen Graphen darstellen.

Eine Menge $\{a_0, a_1\}$, $\{a_1, a_2\}$,..., $\{a_{n-1}, a_n\}$ von Kanten eines Graphen, bei der die Ecken a_0, a_1,..., a_n sämtlich verschieden sind, ist ein **Weg** von a_0 nach a_n. Wenn ein solcher Weg existiert, dann sind a_0 und a_n (durch den Weg) **verbunden**. Ein Graph, in dem je zwei Ecken verbunden sind, ist **zusammenhängend**. Einen Graphen, in dem je zwei Ecken durch genau einen Weg verbunden sind, nennt man in der Graphentheorie einen **Baum**. Abbildung 2.4 zeigt ein Diagramm eines Baumes.

Man verwendet auch solche Bäume zur Modellierung verzweigender Entstehungsprozesse. Man kann wieder die Ecken als „Sprachen“ deuten; eine verbindende Kante gibt dann verwandtschaftliche Nachbarschaft an. Weil ein solcher Baum aber keine Wurzel hat und die Kanten ungerichtet sind, wird die „Richtung“ der Abstammung nicht dargestellt. Es hat deshalb keinen Sinn, ein solches Baumdiagramm von links nach rechts zu zeichnen, wie wir es für Stammbäume eingeführt hatten.

Das Fehlen einer Wurzel ist tatsächlich der einzige ernstzunehmende Unterschied zur Definition 1. Jeder Stammbaum kann ja offenbar als Baum im obigen Sinne gedeutet werden, indem man die Richtungsinformation vergisst. Die Ecken sind die Sprachen, zwei Sprachen sind durch eine Kante verbunden, wenn eine die Vatersprache der anderen ist.

Ebenso leicht kann man aus jedem ungerichteten Baum einen Stammbaum machen, indem man eine (beliebige) Ecke zur Wurzel erklärt und dann für Ecken x, y

(mit $x \neq y$)

$$x < y \quad :\Longleftrightarrow \quad \text{der Weg von der Wurzel nach } x \text{ enthält } y$$

definiert. Diese Relation erfüllte dann die Axiome aus Abschnitt 2.1.

Kapitel 3

Stammbaumrekonstruktion

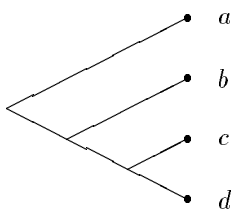
Beim Kopierexperiment, aber auch bei ernsthaften Anwendungen in Biologie und Sprachwissenschaft, sind die Stammbäume nicht explizit bekannt. Die Herausforderung besteht vielmehr darin, die Baumstruktur zu erschließen. Eine besondere Rolle spielen dabei diejenigen Sprachen der Auswahl, die keinen Nachfahren haben. In den meisten Fällen sind dies zugleich die Sprachen, die „bekannt“ sind und einen Namen bekommen; auf ihrer Basis wird die Stammbaumrekonstruktion vorgenommen. Wir wollen diese Menge mit dem Buchstaben B bezeichnen, man darf dabei an die *benannten* Sprachen denken, aber auch an *Blätter* des Baumes[1].

Wie man dazu vorzugehen hat, ob und wie man z.B. aus Merkmalübereinstimmungen oder aus Ähnlichkeiten auf gemeinsame Vorfahren schließen kann, wird später angeschnitten. Zuvor wollen wir mit der schon gewohnten Pingeligkeit festlegen, in welchen Termini dies geschehen soll.

3.1 Stammgruppen

Abstrakt stellt sich die Aufgabe so: Wie beschreibt man einen Wurzelbaum, bei dem nur die Blätter Namen tragen? Es gibt dazu verschiedene einfache Möglichkeiten, die wir kurz skizzieren.

Bei dem nebenstehend abgebildeten Baum ist es z.B. so, daß außer den Blattsprachen a, b, c, d noch drei weitere gemeint sind: eine mit den Nachfahren c und d, eine mit den Nachfahren b, c und d und die Wurzelsprache, die alle anderen als Nachfahren hat. Wir haben damit jeden Verzweigungspunkt durch die Menge seiner Nachfahren unter den Blattsprachen beschrieben; diese Beschreibung funktioniert für beliebige Bäume, in denen es keine Sprachen mit genau einer Tochtersprache gibt. Für die Rekonstruktion sind solche Sprachen unerheblich; im folgenden betrachten wir nur noch **reduzierte** Bäume, das sind solche ohne Sprachen mit genau einer Tochtersprache.



Es sei also $(S, \leq)$ ein Stammbaum und B die Menge seiner Blätter. Zu jeder Sprache $s \in S$ gehört auf naheliegende Weise eine Teilmenge $Sg(s)$ von B, bestehend aus denjenigen Sprachen in B, die von s abstammen. Genauer sei

$$Sg(s) := \{b \in B \mid b \leq s\}.$$

[1] Über die Abstammung der Sprachen gibt es historische Informationen. Deshalb ist die Einschränkung, nur die Blätter des Stammbaumes (also nur die rezenten Sprachen) als bekannt anzunehmen und die übrigen Sprachen als rekonstruierte Verzweigungspunkte zu verstehen, zu restriktiv. In einer späteren Fassung muß dies verbessert werden; zunächst bleiben wir aber der Übersichtlichkeit halber bei dieser Voraussetzung.

Teilmengen der Form $Sg(s)$ nennen wir **Stammgruppen** (der Biologe spricht von **monophyletischen Gruppen**[2]). Für die Menge aller Stammgruppen des jeweils betrachteten Stammbaumes benutzen wir das Symbol $\mathcal{S}$.

Satz 1 *Eine Menge $\mathcal{S}$ von Teilmengen einer endlichen Menge B ist genau dann die Menge aller Stammgruppen eines Baumes (mit Blättermenge B), wenn folgende Bedingungen erfüllt sind:*

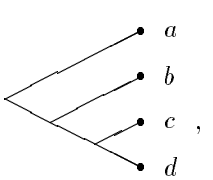
1. $B \in \mathcal{S}$, $\emptyset \notin \mathcal{S}$,
2. $\{b\} \in \mathcal{S}$ *für alle* $b \in B$,
3. *Für* $G, H \in \mathcal{S}$ *gilt stets* $G \cap H = \emptyset$, $G \subseteq H$ *oder* $H \subseteq G$.

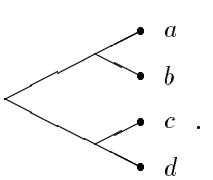
Beweis Daß die Menge der Stammgruppen eines Baumes die ersten beiden Bedingungen erfüllt, ergibt sich direkt aus der Definition. Sind G und H Stammgruppen, dann sind $g := \bigvee G$ und $h := \bigvee H$ Elemente des Baumes mit $G = Sg(g)$ und $H = Sg(h)$. Wenn nun $G \cap H \neq \emptyset$ gilt, wenn also z.B. a ein Element von $G \cap H$ ist, dann gilt $a < g$ und $a < h$, nach Axiom 6 also $g \leq h$ oder $h \leq g$ und damit $G \subseteq H$ oder $H \subseteq G$.

Ist umgekehrt $\mathcal{S}$ ein Mengensystem, welches die Bedingungen des Hilfssatzes erfüllt, dann ist $(\mathcal{S}, \subseteq)$ eine endliche geordnete Menge mit größtem Element B, und die Bedingung 3 des Hilfssatzes stellt sicher, daß das Axiom 6 erfüllt ist. $(\mathcal{S}, \subseteq)$ ist also ein Wurzelbaum, dessen Blätter gerade die minimalen Mengen in $\mathcal{S}$, also die Mengen der Form $\{b\}, b \in B$ sind. Benennen wir diese um, indem wir die Mengenklammern weglassen, so erhalten wir einen Baum mit Blättermenge B und Stammgruppenmenge $\mathcal{S}$. □

Eine Mengenfamilie, die die im Hilfssatz angegebenen Bedingungen erfüllt, nennen wir auch einen **Mengenbaum**. Die Stammgruppen enthalten die Information über die Struktur des Stammbaumes vollständig, außer daß sich die Namen der Sprachen, die nicht zu B gehören, aus ihnen nicht ablesen lassen. Zwei reduzierte Stammbäume mit den gleichen Stammgruppen sind *isomorph* (=strukturgleich), und $(S, \leq)$ ist isomorph zu $(\mathcal{S}, \subseteq)$, falls $(S, \leq)$ reduziert ist. Die Menge B und die Mengen $\{b\}$ sind nach dem Hilfssatz stets Stammgruppen; aufgeschrieben werden deshalb meist nur die übrigen, die **echten** Stammgruppen.

Wir zeigen an einem kleinen Beispiel, wie die Beschreibung mittels der Stammgruppen genutzt werden kann, wenn man fehlerhafte Beschreibungen zu bearbeiten hat. Die Information über die Abstammung von vier Sprachen a, b, c, d möge in zwei Versionen vorliegen,

einmal als 

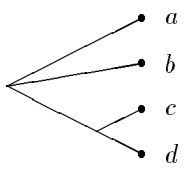
, sowie als 

.

Die echten Stammgruppen sind im ersten Fall $\{c, d\}$, $\{b, c, d\}$, im zweiten Fall $\{a, b\}$ und $\{c, d\}$. Gemeinsam ist also die echte Stammgruppe $\{c, d\}$. Tatsächlich gibt es

[2] Zum (abweichenden) Gebrauch der Bezeichnung „Stammgruppe“ in der Biologie vergl. §12.4 in [18].

einen Baum, der nur diese echte Stammgruppe hat, nämlich



Wir können diesen Baum also als die gemeinsame Vergröberung der beiden angegebenen Bäume ansehen. Dieser Trick funktioniert allgemein:[3]

Korollar 1 *Ist $\mathcal{S}_1$ die Menge der Stammgruppen eines Baumes mit Blättermenge B und $\mathcal{S}_2$ die Menge der Stammgruppen eines zweiten Baumes mit der gleichen Blättermenge, dann ist auch $\mathcal{S}_1 \cap \mathcal{S}_2$ die Menge der Stammgruppen eines solchen Baumes.*

Wir haben im Beispiel gesehen, daß die Mengen der Stammgruppen zweier Bäume ineinander enthalten sein können. Ist dies der Fall (und haben beide Bäume die gleiche Blättermenge), dann nennen wir den Baum mit weniger Stammgruppen **gröber** und den anderen **feiner**. Es gibt immer einen gröbsten Baum, nämlich den, der überhaupt keine echten Stammgruppen hat. Ein Baum ist genau dann **nicht verfeinerbar**, was besagen soll, daß seine Stammgruppenmenge nicht erweitert werden kann, wenn jeder Verzweigungspunkt genau zwei Töchter hat.

Eine weitere Folgerung schließt sich an:

Korollar 2 *Ist $\mathcal{S}$ die Menge aller Stammgruppen eines Baumes mit Blättermenge B und ist T eine Teilmenge von B, dann ist*

$$\{S \cap T \mid S \in \mathcal{S}, S \cap T \neq \emptyset\}$$

die Menge der Stammgruppen eines Baumes mit Blättermenge T.

In der von W. Hennig begründeten Form der biologischen Phylogenetik, der **Cladistik**, werden die Stammgruppen zur Rekonstruktion benutzt. Es wird angenommen, daß jeder Verzweigungspunkt im Stammbaum genau zwei Töchter hat. Hauptschritt der Rekonstruktion ist es dann, zu jeder Stammgruppe die **Schwestergruppe** zu bestimmen, also diejenige, die den gleichen Vater hat.

3.2 Das zweite Kopierexperiment

Wir illustrieren die im vorigen Abschnitt erarbeiteten Begriffe an einem Beispiel. Unser Erkenntnisfortschritt ist allerdings noch viel zu klein, als dass wir uns an ernstzunehmende sprachwissenschaftliche Daten heranwagen könnten. Nicht einmal die Daten des ersten Fotokopierexperimentes lassen sich damit interpretieren. Wir diskutieren deshalb ein ähnliches Experiment, das aber im Vergleich zu dem in Abschnitt 1.2 stark vereinfacht und dadurch leicht zu entschlüsseln ist. Der Unterschied besteht darin, dass diesmal Wörter nicht durch andere *ersetzt* werden, sondern neue Wörter zu den alten *hinzugefügt* werden.

Das Experiment beginnt diesmal mit einem Blatt Papier, welches ein Tabellenschema mit sechs Feldern zeigt:

[3] In der Mathematik müssen alle Lehrsätze bewiesen oder durch den Verweis auf einen Beweis gerechtfertigt werden. Manchmal hat ein Satz Konsequenzen, die ganz offensichtlich sind, bei denen es sich dennoch lohnt, sie auszusprechen. Dann spricht man von einem Korollar („Krönchen“) zum Lehrsatz.

Der Experimentator kopiert dieses Blatt und gibt Original und Kopie ins Publikum mit der Bitte, jeweils ein oder zwei frei gewählte Wörter einzutragen. Ein zurücklaufendes Blatt könnte also beispielsweise folgendermaßen aussehen:

	Gebet
krank	

Die rücklaufenden Blätter werden nun ihrerseits kopiert und wieder mit der gleichen Bitte ins Publikum gegeben, und so fort. Das Experiment wurde abgebrochen, als zehn weitgehend ausgefüllte Blätter abgegeben waren; die bei Abbruch des Experimentes noch im Publikum befindlichen Blätter wurden ignoriert. Das Ergebnis ist in Abbildung 3.1 gezeigt.

1

Birke	Jubel
Sinus	
jeder	Grube

2

famos	Wurst
Staub	
Paris	Chili

3

neben	Jubel
Iltis	Onkel
gehen	Roman

4

hurra	Jubel
Sinus	Katze
jeder	Spiel

5

Staub	Jubel
Iltis	Papst
	Maske

6

Scham	Kraut
Quark	Aroma
Paris	Chrom

7

	Kraut
	Tasse
Paris	Hilfe

8

famos	Schau
Staub	Humus
Paris	Niete

9

Lehre	Jubel
Sinus	Abtei
jeder	Gnade

10

famos	Wurst
Staub	Creme
Paris	Traum

Abbildung 3.1: Die Daten eines vereinfachten Kopierexperimentes. In jedem Schritt wurde das jeweilige Formular kopiert und ein oder zwei Wörter zusätzlich eingetragen. Weil keine Wörter ersetzt wurden, ist der Stammbaum leicht zu rekonstruieren.

Es ist wiederum plausibel, dass der Ablauf des Experimentes durch einen Stammbaum beschrieben werden kann, dessen Blätter die zurückgekehrten Tabellen sind. Es ist nicht schwierig, aus diesen Daten den Stammbaum zu rekonstruieren. Der Leser möge dies zunächst einmal ohne Anleitung versuchen.

Grundlage unserer Rekonstruktion ist das mehrfache Auftreten einzelner Wörter. Wir dürfen annehmen, dass diese Dopplungen durch das Kopieren zustandegekommen sind, können dabei allerdings nicht ganz sicher sein. Es könnte ja auch ein Wort zweimal eingetragen worden sein. Allerdings scheint dies wenig wahrscheinlich. (Im vorliegenden Datensatz hat es eine solche Wortverdopplung tatsächlich gegeben, wie

	1	2	3	4	5	6	7	8	9	10
Jubel	×		×	×	×				×	
Sinus, jeder	×			×					×	
famos, Staub(m)		×						×		×
Wurst		×								×
Paris		×				×	×	×		×
Iltis			×		×					
Kraut						×	×			

Abbildung 3.2: Die Wörter, die in Abbildung 3.1 mehrfach auftreten.

man rasch entdeckt: Das Wort „Staub“ tritt in Blatt 2 und Blatt 5 auf, aber nicht an der gleichen Stelle. Das kann natürlich nicht durch Kopieren entstanden sein. Wir müssen deshalb „Staub (oben)“ und „Staub (mitte)“ wie verschiedene Wörter behandeln.)

Die einfachste Rekonstruktionsmethode ist hier die Bestimmung der Stammgruppen. Die trivialen Stammgruppen bestehen aus genau einem oder aus allen Blättern. Jede nichttriviale Stammgruppe besteht aus den „Nachkommen“ einer Zwischenstufe. Weil in unserem Experiment nichts gelöscht wurde, erhalten die Nachkommen alle Merkmale ihrer Vorgänger. Der gemeinsame Ahn einer nichttrivialen Stammgruppe enthält mindestens ein neues Merkmal; dieses ist allen seinen Nachkommen gemeinsam und kommt in den anderen Blättern nicht vor.

Dies lässt sich sehr einfach in eine Handlung umsetzen: Wir notieren, welche Wörter in mehr als einem Formular auftreten. Dies ist in Abbildung 3.2 angegeben.

Die Zeilen dieser Tabelle geben die nichttrivialen Stammgruppen an. Man erkennt, dass diese Mengen die in Satz 1 angegebenen Eigenschaften haben, die Daten sind also wohl in Ordnung. Fügt man die trivialen Stammgruppen hinzu, so erkennt man sogleich die Struktur des Mengenbaumes:

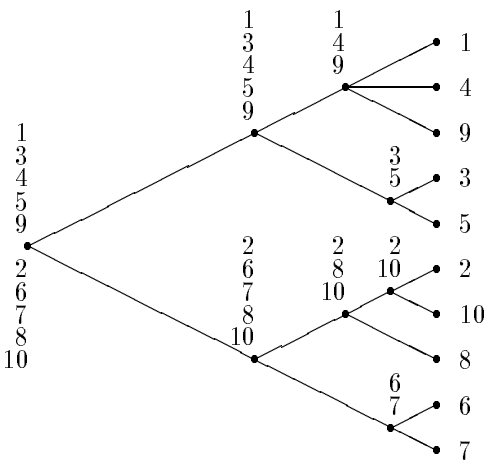


Abbildung 3.3: Der Mengenbaum entsteht aus den Stammgruppen.

Damit ist das Rekonstruktionsvorhaben gelungen. Anhand der Originaldaten kann man nun noch die jeweiligen Veränderungen an den Kanten eintragen, was allerdings zu Platzproblemen beim Beschriften führt. Damit ist die gesamte Entwicklung mit allen Zwischenstufen bestimmt.

An dem in Abbildung 3.4 wiedergegebenen Ergebnis lässt sich noch einiges ab-

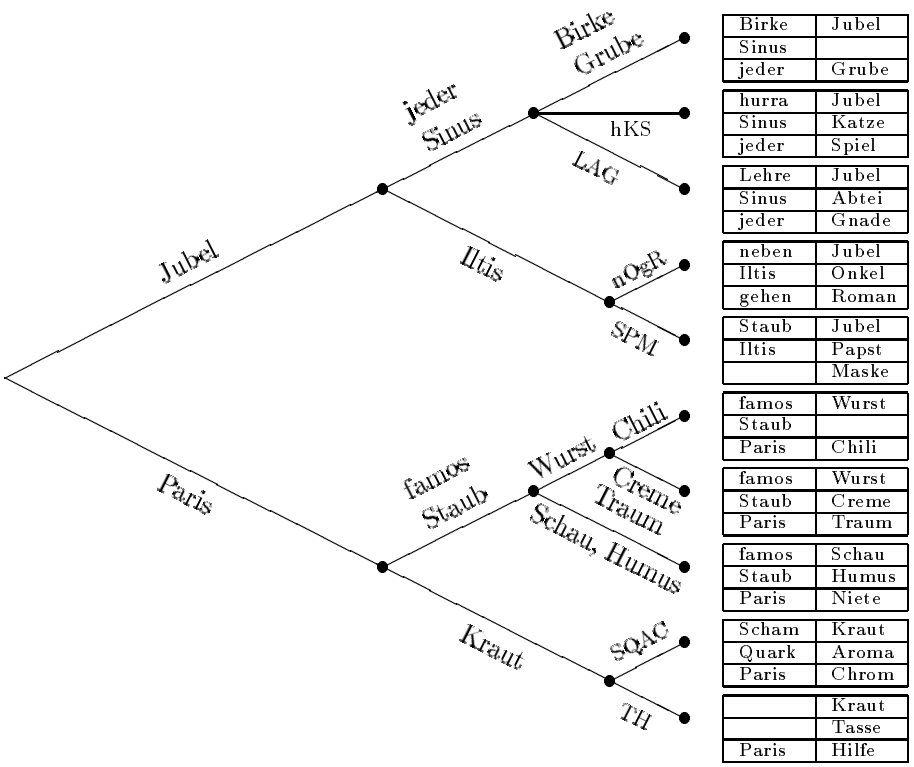


Abbildung 3.4: Die Rekonstruktion der Entwicklung beim zweiten Kopierexperiment. Die Abkürzungen stehen für die Wörter in der jeweiligen Tabelle, die an dem betreffenden Pfad anderweitig nicht eingetragen sind.

lesen. Offenbar ist während des Experimentes ein Fehler passiert, denn eine Dreifachverzweigung (Trifurkation), wie sie im Ergebnis auftritt, ist eigentlich nicht mit dem geplanten Versuchsablauf im Einklang. Offenbar ist hier entweder ein Zettel zweimal kopiert worden oder, was auf das gleiche hinausläuft, ohne Änderungen zurückgegeben (und dann erneut kopiert) worden. Auf den ersten Blick scheint es auch fehlerhaft, dass an einzelnen Kanten bis zu vier Änderungen vermerkt sind; es sollten doch jeweils nur ein oder zwei Wörter eingetragen werden. Das erklärt sich aber daraus, dass nicht alle Ergebnisblätter in die Auswertung einbezogen wurden. Dadurch kommt es vor, dass Zwischenstufen nur einen Nachkommen in die Auswertung eingebracht haben. Solche Zwischenstufen werden bei der Rekonstruktion nicht berücksichtigt. Es kann deshalb sein, dass eine Kante für mehr als einen Änderungsvorgang steht. Deshalb können auch mehrere Wörter an einer Kante stehen.

3.3 Versus-Relationen und Implikationen

Die Beschreibung mit Hilfe der Stammgruppen ist präzise, aber oft unhandlich. Man wünscht sich eine mehr lokale Beschreibung also eine, bei der über einzelne Sprachen geredet wird, also mit Aussagen wie

> *a und b haben einen gemeinsamen Vorfahren*

gearbeitet wird. Ganz so einfach kann es allerdings nicht gehen, denn aufgrund unserer Modellannahmen ist diese Aussage für je zwei der betrachteten Sprachen richtig und ihre Negation damit stets falsch. Geeignetes Sprachmittel ist, wie schon der Biologe W. Hennig herausstellte, eine Aussage über drei Sprachen:

> *a und b haben einen gemeinsamen Vorfahren, der nicht Vorfahr von c ist.*

Weil dies etwas umständlich auszusprechen ist, führen wir dafür die Kurzschreibweise

$$ab \,|\, c$$

ein, die wir gern „a, b **versus** c“ lesen. Aus jedem Stammbaum erhalten wir so eine dreistellige Relation auf B, die wir die **Versus-Relation** des Stammbaumes nennen. $ab \,|\, c$ und $ba \,|\, c$ sollen gleichbedeutend sein und nur für $a \neq c$ benutzt werden. $aa \,|\, b$ soll für $a \neq b$ stets gelten, wird aber gewöhnlich nicht extra erwähnt[4].

Wir führen auch eine Abkürzung für die Negation ein: Um auszudrücken, daß $ab \,|\, c$ *nicht* gilt, schreiben wir $\neg(ab \,|\, c)$ oder $ab \to c$. Die zweite Schreibweise wird verständlicher, wenn man auf die früher definierte Operation $\vee$ zurückgreift. Man erhält dann als gleichwertige Definition

$$ab \,|\, c \qquad \Longleftrightarrow \qquad c \not\leq a \vee b$$

und

$$ab \to c \qquad \Longleftrightarrow \qquad c \leq a \vee b.$$

Die Versus-Relation eines Stammbaumes enthält alle wesentlichen Strukturinformationen. Um dies nachzuweisen, zeigen wir, daß sich aus der Versus-Relation die Stammgruppen rekonstruieren lassen (und damit auch der reduzierte Stammbaum).

Hilfssatz 2 *$A \subseteq B$ mit $A \neq \emptyset$ ist genau dann eine Stammgruppe, wenn $ab \,|\, c$ für alle $a, b \in A$ und alle $c \notin A$ gilt.*

[4] Die binäre Relation $aa \,|\, b$ drückt aus, dass b nicht von a abstammt. Sie spielt eine Rolle, wenn auch Nichtblattsprachen einbezogen werden.

Beweis Wenn A eine Stammgruppe ist, also wenn $A = Bl(s)$ für ein Element $s \in S$ gilt, dann gilt für $a, b \in A$ und $c \notin A$ stets $a \vee b \leq s$ und $c \not\leq s$, was zusammen $c \not\leq a \vee b$, also $ab \,|\, c$ ergibt.

Für die Umkehrung setze $s := \bigvee A$. Wir behaupten $A = Bl(s)$. Sicherlich gilt $A \subseteq Bl(s)$. Ist $c \in Bl(s)$, gilt also $c \leq \bigvee A$, dann findet man nach Hilfssatz 1 Elemente a und b in A mit $c \leq a \vee b = \bigvee A$. Es gilt also dann $\neg(ab \,|\, c)$, und deshalb $c \in A$. □

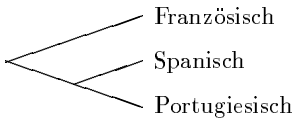
Beispiel 1 Eine Versus-Relation soll für die Sprachen

Okzitanisch, Katalanisch, Französisch, Spanisch, Portugiesisch, Gallego

angegeben werden. Wir entwerfen dazu ein Formular, welches für jede mögliche Auswahl von drei der sechs Sprachen ein Feld enthält, das nach folgendem Prinzip auszufüllen ist:

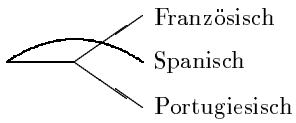
	Französisch
×	Spanisch
×	Portugiesisch

steht für



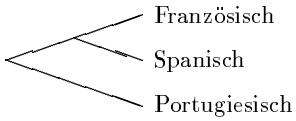
×	Französisch
	Spanisch
×	Portugiesisch

steht für



×	Französisch
×	Spanisch
	Portugiesisch

steht für



$ab \,|\, c$ wird also angegeben, indem man bei a und b ein Kreuz macht, bei c hingegen keins. Abbildung 3.5 zeigt ein ausgefülltes Formular, Abbildung 3.6 den zugehörigen Baum. Man beachte, daß wir nicht behaupten, dieser Baum gebe die Entstehungsgeschichte richtig[5] wieder, sondern nur, daß er die oben angegebenen Daten getreu abbildet. Man vergewissere sich, daß dies stimmt!

Der im Beipiel vorgestellte Weg, wie ein Stammbaum aus einer Versus-Relation gewonnen werden kann, wirft mehrere Fragen auf:

- *Woher kommen die Daten?* Darüber soll später gesprochen werden; im Moment geht es darum, daß diese Form der Darstellung den Stammbaum beschreibt. Diese Angaben können z.B. einfach den momentanen Kenntnisstand eines Sprachwissenschaftlers wiedergeben.

- *Ist die Darstellung nicht zu kompliziert?* Man kann leicht Abkürzungen angeben, beispielsweise durch Zusammenfassung von Sprachen. An dieser Stelle geht es uns aber darum, das Prinzip der Schreibweise offenzulegen.

- *Wie kommt man von der Tabelle zum Baum?* Wir geben dazu einen Algorithmus an, also ein Verfahren, welches man ohne Weiteres abarbeiten kann und das sich auch leicht in ein Computerprogramm gießen lässt. Weil wir aber

[5] Wir haben die Daten in unseren Beispielen übrigens nicht erfunden. Sie beruhen vielmehr auf Daten, die uns Sprachwissenschaftler überlassen haben, sind also Experteneinschätzungen. Allerdings sind diese Einschätzungen teilweise sehr spontan formuliert worden.

	Okzitanisch
	Katalanisch
	Französisch

×	Okzitanisch
×	Katalanisch
	Spanisch

×	Okzitanisch
×	Katalanisch
	Portugiesisch

×	Okzitanisch
×	Katalanisch
	Gallego

×	Okzitanisch
×	Französisch
	Spanisch

×	Okzitanisch
×	Französisch
	Portugiesisch

×	Okzitanisch
×	Französisch
	Gallego

	Okzitanisch
×	Spanisch
×	Portugiesisch

	Okzitanisch
×	Spanisch
×	Gallego

	Okzitanisch
×	Portugiesisch
×	Gallego

×	Katalanisch
×	Französisch
	Spanisch

×	Katalanisch
×	Französisch
	Portugiesisch

×	Katalanisch
×	Französisch
	Gallego

	Katalanisch
×	Spanisch
×	Portugiesisch

	Katalanisch
×	Spanisch
×	Gallego

	Katalanisch
×	Portugiesisch
×	Gallego

	Französisch
×	Spanisch
×	Portugiesisch

	Französisch
×	Spanisch
×	Gallego

	Französisch
×	Portugiesisch
×	Gallego

×	Spanisch
×	Portugiesisch
	Gallego

Abbildung 3.5: Angabe einer Versus-Relation

schon sehr bald ein viel besseres Verfahren vorstellen werden, verzichten wir auf eine Begründung.

Wir gehen von einem Datensatz aus, wie er in Abbildung 3.5 angegeben ist, lesen ihn als Versus-Relation und konstruieren daraus das Baumdiagramm „von rechts nach links“, von den Blättern zur Wurzel, indem wir Teildiagramme erzeugen, die schrittweise wachsen.

Wir bezeichnen die Menge der betrachteten Sprachen wieder mit B und die zu konstruierende Menge der Stammgruppen mit $\mathcal{S}$.

1. Setze $C := B$, $\mathcal{S} := \{\{b\} \mid b \in B\}$.
2. Falls $|C| \leq 1$: Ende.
3. $W := C$.
4. Suche Elemente $a, b \in W$, $a \neq b$, so dass für alle $x \in W$ mit $a \neq x \neq b$ gilt: $ax \to b$ und $bx \to a$.
5. Setze $T := \{x \in C \setminus \{a, b\} \mid ab \to x\}$. Wenn $|T| > 1$, dann setze $W := T$ und gehe nach 4.
6. Setze
$$G_{a,b} := \{x \in B \mid ab \to x\},$$
ersetze C durch $C \setminus G_{a,b} \cup \{a\}$ und $\mathcal{S}$ durch $\mathcal{S} \cup G_{a,b}$.
7. Gehe zu 2.

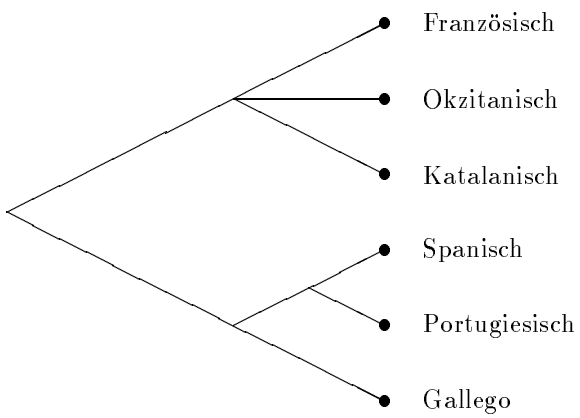


Abbildung 3.6: Der aus dem Datensatz konstruierte Baum.

3.4 Charakterisierung der Versus-Relationen.

Das oben vorgestellte Verfahren, den Baum aus der Versus-Relation zu berechnen, setzt voraus, daß die Relation vollständig und korrekt vorliegt. Es ist allerdings nicht gesagt, daß der Algorithmus nur bei fehlerfreien Daten funktioniert. Will man nachprüfen, ob ein solches Datenblatt konsistent ausgefüllt ist, so muß man nach dem bisher gesagten zunächst den Stammbaum konstruieren und dann die Angaben damit vergleichen. Das ist umständlich und wenig instruktiv. Deshalb geben wir nun eine einfache Charakterisierung der Versus-Relationen an.

Satz 2 *Eine dreistellige Relation $ab\,|\,c$ auf einer Menge B ist genau dann die Versus-Relation eines Baumes, wenn sie die folgenden Eigenschaften hat:*

1. *$aa\,|\,b$ gilt für alle a, b mit $a \neq b$.*
2. *Aus $ab\,|\,c$ folgt $ba\,|\,c$.*
3. *Aus $ab\,|\,c$ folgt $\neg(ac\,|\,b)$.*
4. *Aus $ab\,|\,c$ folgt $ab\,|\,d$ oder $ad\,|\,c$.*

Eigenschaft 4 nennt man auch die *Vier-Punkte-Bedingung* für Stammbäume.
Beweis Daß eine Versus-Relation die ersten drei Eigenschaften hat, war schon erwähnt. Die vierte Eigenschaft ergibt sich so: aus $ab\,|\,c$, also $c \not\leq a \vee b$, und $\neg(ab\,|\,d)$, also $d \leq a \vee b$, folgt $a \vee d \leq a \vee b$ und folglich $c \not\leq a \vee d$, was $ad\,|\,c$ bedeutet.

Zum Beweis der Umkehrung definiere $\mathcal{S}$ als die Menge aller nichtleeren Teilmengen $G \subseteq B$ mit der Eigenschaft, daß $ab\,|\,c$ für alle $a, b \in G$ und alle $c \notin G$ gilt. Wir zeigen, daß diese Mengenfamilie die Bedingungen aus Satz 2 erfüllt. Die ersten beiden Bedingungen sind offensichtlich erfüllt. Gäbe es Mengen G und H in $\mathcal{S}$, die die dritte Bedingung verletzen, dann müßte es Elemente $a, b, c \in B$ geben mit $a \in G \setminus H$, $b \in G \cap H$, $c \in H \setminus G$. Das hätte aber $ab\,|\,c$ und $bc\,|\,a$ zu Folge, was Bedingung 3 des Satzes widerspricht. $\mathcal{S}$ ist also die Menge der Stammgruppen eines Baumes. Zu zeigen ist nun noch, daß die Ausgangsrelation die nach Hilfssatz 2 zu diesem Baum gehörende Versus-Relation ist; dazu müssen wir nachweisen, daß $ab\,|\,c$ tatsächlich nur dann gilt, wenn es eine Menge $G \subseteq \mathcal{S}$ gibt mit $a, b \in G, c \notin G$. Dazu zeigen wir, daß $G := \{x \mid ax|c\}$ zu $\mathcal{S}$ gehört, und dies wiederum erfordert den Nachweis, daß für alle $u, v \in G$ und $w \notin G$ die Bedingung $uv\,|\,w$ gilt. Wenn $au\,|\,c$

gilt, dann muß nach Bedingung 4 $au \mid w$ oder $aw \mid c$ gelten; im Falle $w \notin G$ kommt also nur $au \mid w$ in Frage. Hat man also $u, v \in G$ und $w \notin G$, so ergibt sich $au \mid w$ und $av \mid w$, und unser Beweis ist komplett, wenn wir noch folgendes beweisen:

Aus $au \mid w$ und $av \mid w$ folgt $uv \mid w$.

Das ergibt sich mit den Bedingungen 2, 3 und 4: aus $au \mid w$ erhält man $ua \mid w$ und daraus $ua \mid v$ oder $uv \mid w$; ebenso folgt aus $av \mid w$ dann $va \mid u$ oder $uv \mid w$. Da aber $ua \mid v$ und $va \mid u$ nicht zugleich richtig sein können, ist $uv \mid w$ erzwungen. □

Korollar 3 *Eine dreistellige Relation $ab \to c$ auf einer Menge B ist genau dann das Komplement der Versus-Relation eines Baumes, wenn sie die folgenden Eigenschaften hat:*

1. *Aus $aa \to b$ folgt $a = b$.*
2. *Aus $ab \to c$ folgt $ba \to c$.*
3. *Aus $ab \not\to c$ folgt $ac \to b$.*
4. *Aus $ab \to d$ und $ad \to c$ folgt $ab \to c$.*

Der Satz hat, trotz seines einfachen Beweises, einige bemerkenswerte Konsequenzen. Er zeigt unter anderem:

- Jede Einschränkung einer Versus-Relation auf eine Teilmenge ist ebenfalls eine Versus-Relation.
- Versus-Relation zu sein ist eine lokale Eigenschaft: wenn eine Relation nicht Versus-Relation ist, dann gibt es schon eine vierelementige Teilmenge der Grundmenge, so daß die Einschränkung auf diese Teilmenge nicht Versus-Relation ist.
- Partielle Informationen über einen Stammbaum können dadurch vereinigt werden, daß man die zugehörigen Versus-Relationen vereinigt.

Beispiel 2 Wiederum soll eine Versus-Relation für die Sprachen

Okzitanisch, **K**atalanisch, **F**ranzösisch, **S**panisch, **P**ortugiesisch, **G**allego

angegeben werden. Das entsprechende Formular wurde wie in Abbildung 3.7 ausgefüllt. Es gelingt nicht, diese Angaben in einen Stammbaum einzupassen, denn die Bedingung 4 aus Satz 2 ist verletzt, und zwar gleich mehrfach. Die (von Hand mühsame) Überprüfung aller Möglichkeiten ergibt, daß die Bedingung für fünf Auswahlen von jeweils vier Sprachen verletzt ist, nämlich für:

Okzitanisch – Katalanisch – Französisch – Spanisch
Okzitanisch – Katalanisch – Französisch – Portugiesisch
Okzitanisch – Katalanisch – Französisch – Gallego
Okzitanisch – Katalanisch – Spanisch – Portugiesisch
Okzitanisch – Katalanisch – Spanisch – Gallego.

Vermutlich hat hier Unsicherheit bezüglich der Einordnung des Okzitanischen zusammen mit dem Katalanischen bestanden.

×	Okzitanisch
×	Katalanisch
	Französisch

×	Okzitanisch
×	Katalanisch
	Spanisch

×	Okzitanisch
×	Katalanisch
	Portugiesisch

×	Okzitanisch
×	Katalanisch
	Gallego

×	Okzitanisch
×	Französisch
	Spanisch

×	Okzitanisch
×	Französisch
	Portugiesisch

×	Okzitanisch
×	Französisch
	Gallego

	Okzitanisch
×	Spanisch
×	Portugiesisch

	Okzitanisch
×	Spanisch
×	Gallego

	Okzitanisch
×	Portugiesisch
×	Gallego

×	Katalanisch
	Französisch
×	Spanisch

×	Katalanisch
	Französisch
×	Portugiesisch

×	Katalanisch
	Französisch
×	Gallego

×	Katalanisch
×	Spanisch
	Portugiesisch

×	Katalanisch
×	Spanisch
	Gallego

	Katalanisch
×	Portugiesisch
×	Gallego

	Französisch
×	Spanisch
×	Portugiesisch

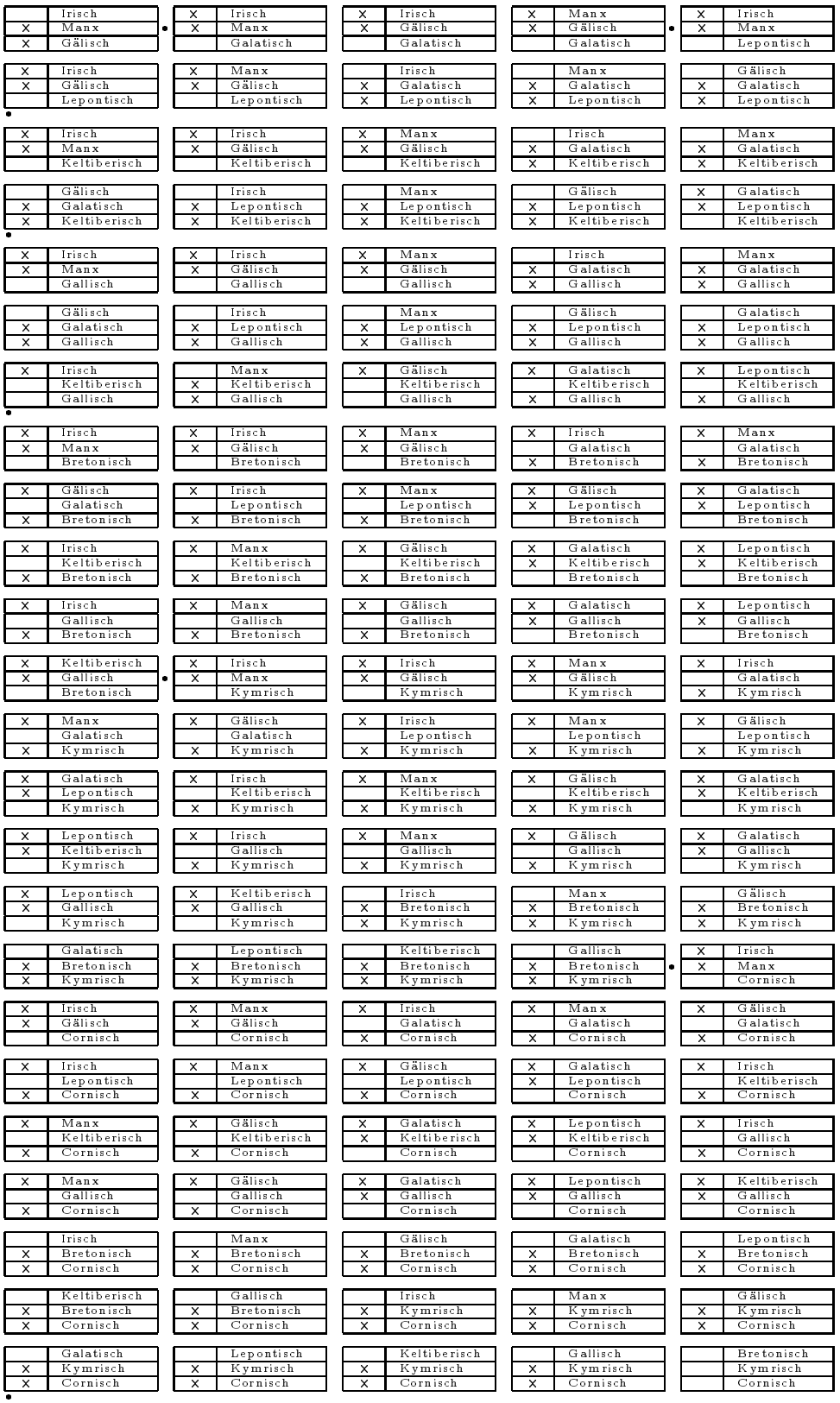
	Französisch
×	Spanisch
×	Gallego

	Französisch
×	Portugiesisch
×	Gallego

	Spanisch
×	Portugiesisch
×	Gallego

Abbildung 3.7: Die Bedingungen für eine Versus-Relation sind nicht erfüllt.

Übungsaufgabe 1 Von der Sprachwissenschaftlerin Katharina Kupfer (jetzt Universität Frankfurt) erhielten wir ein folgendermaßen ausgefülltes Datenblatt:



×	Irisch
×	Manx
	Kumbrisch

×	Irisch
×	Gälisch
	Kumbrisch

×	Manx
×	Gälisch
	Kumbrisch

×	Irisch
	Galatisch
×	Kumbrisch

×	Manx
	Galatisch
×	Kumbrisch

×	Gälisch
	Galatisch
×	Kumbrisch

×	Irisch
	Lepontisch
×	Kumbrisch

×	Manx
	Lepontisch
×	Kumbrisch

×	Gälisch
	Lepontisch
×	Kumbrisch

×	Galatisch
×	Lepontisch
	Kumbrisch

×	Irisch
	Keltiberisch
×	Kumbrisch

×	Manx
	Keltiberisch
×	Kumbrisch

×	Gälisch
	Keltiberisch
×	Kumbrisch

×	Galatisch
×	Keltiberisch
	Kumbrisch

×	Lepontisch
×	Keltiberisch
	Kumbrisch

×	Irisch
	Gallisch
×	Kumbrisch

×	Manx
	Gallisch
×	Kumbrisch

×	Gälisch
	Gallisch
×	Kumbrisch

×	Galatisch
×	Gallisch
	Kumbrisch

×	Lepontisch
×	Gallisch
	Kumbrisch

×	Keltiberisch
×	Gallisch
	Kumbrisch

	Irisch
×	Bretonisch
×	Kumbrisch

	Manx
×	Bretonisch
×	Kumbrisch

	Gälisch
×	Bretonisch
×	Kumbrisch

	Galatisch
×	Bretonisch
×	Kumbrisch

	Lepontisch
×	Bretonisch
×	Kumbrisch

	Keltiberisch
×	Bretonisch
×	Kumbrisch

	Gallisch
×	Bretonisch
×	Kumbrisch

	Irisch
×	Kymrisch
×	Kumbrisch

	Manx
×	Kymrisch
×	Kumbrisch

	Gälisch
×	Kymrisch
×	Kumbrisch

	Galatisch
×	Kymrisch
×	Kumbrisch

	Lepontisch
×	Kymrisch
×	Kumbrisch

	Keltiberisch
×	Kymrisch
×	Kumbrisch

	Gallisch
×	Kymrisch
×	Kumbrisch

	Bretonisch
×	Kymrisch
×	Kumbrisch

	Irisch
×	Cornisch
×	Kumbrisch

	Manx
×	Cornisch
×	Kumbrisch

	Gälisch
×	Cornisch
×	Kumbrisch

	Galatisch
×	Cornisch
×	Kumbrisch

	Lepontisch
×	Cornisch
×	Kumbrisch

	Keltiberisch
×	Cornisch
×	Kumbrisch

	Gallisch
×	Cornisch
×	Kumbrisch

	Bretonisch
×	Cornisch
×	Kumbrisch

	Kymrisch
	Cornisch
	Kumbrisch

Bestimmen Sie den zugehörigen Stammbaum.

Literaturverzeichnis

[1] A.V. Aho, Y. Sagiv, T.G. Szymanski and J.D. Ullman. *Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions.* Siam Journal of Computing 10:3 (1981), 405–421.

[2] G. Altmann und W. Lehfeldt: *Allgemeine Sprachtypologie. Prinzipien und Meßverfahren.*

[3] Hans-Jürgen Bandelt und Andreas Dress. *Reconstructing the shape of a tree from observed dissimilarity data.* Advances in Applied Mathematics 7 (1986), 309–343.

[4] P. Buneman. *The recovery of trees from measures of dissimilarity.* In: F.R. Hodson, D.G. Kendall, and P. Tautu, eds., *Mathematics in the archaeological and historical sciences.* Edinburgh University Press 1971, 387–395.

[5] H. Colonius und H.H. Schulze. *Tree structures for proximity data.* British Journal Math. Statistical Psychology 34 (1981), 167–180.

[6] Isidore Dyen, Joseph B. Kruskal and Paul Black. *An Indoeuropean classification: a lexicostatistical experiment.* Transactions of the American Philosophical Society 82:5 (1992). ISBN 0-87169-825-0.

[7] I. Fodor. *The validity of Glottochronology on the Basis of the Slavonic Languages.* Studia Slavica 7 (1961), 295–346.

[8] Harald Haarmann. *"Basic" Vocabulary and Language Contacts: The Disillusion of Glottochronology.* Indogermanische Forschungen 95(1990), 1–37.

[9] B. Heine, H. Hoff, R. Vossen. *Neuere Ergebnisse zur Territorialgeschichte der Bantu.* In: W.J.G. Möhling, F. Rottland, B. Heine. *Zur Sprachgeschichte und Ethnohistprie in Afrika: Neue Beiträge afrikanistischer Forschungen.* Berlin 1977, 57–72.

[10] Willi Hennig. *Phylogenetische Systematik.* Pareys Studientexte 34, 1982. ISBN 3-489-60934-4.

[11] Joseph H. Greenberg. *Language in the Americas.* Stanford University Press 1987.

[12] April M. S. McMahon and Robert McMahon. *Linguistics, genetics and archaeology: internal and external evidence in the Amerind controversy.* Transactions of the Philological Society 93:2 (1995), 125–225.

[13] Johanna Nichols. *Linguistic diversity and the first settlement of the new world.* Language 66 (1990), 475–521.

[14] C. Rabin. *Lexicostatistics and the Internal Divisions of Semitic.* In: J. Bynon, T. Bynon (eds.) *Hamito-Semitica.* The Hague 1977, 85–99.

[15] J.A. Rea. *Concerning the Validity of Lexicostatistics.* International Journal of American Linguistics 24 (1958), 145–150.

[16] J.M.S. Simões-Pereira. *A note on the tree realizability of a distance matrix.* J. Combinatorial Theory 6 (1969), 303–310.

[17] Michael Steel. *The complexity of reconstructing trees from qualitative characters and subtrees.* Journal of Classification 9 (1992), 91–116.

[18] Walter Sudhaus und Klaus Rehfeld. *Einführung in Phylogenetik und Systematik.* Gustav Fischer Verlag 1992. ISBN 3-437-20475-0.

[19] Morris Swadesh. *Salish internal relationship.* International Journal of American Linguistics 16, 157–167.

[20] Johann Tischler. *Glottochronologie und Lexikostatistik.* Dissertation Universität Innsbruck 1973.