Theoretische und praktische Aspekte automatischer Sequenzanalysen in der historischen Linguistik

Johann-Mattis List

15. April 2010

Im Folgenden soll ein Überblick über theoretische und praktische Fragen zur Implementierung von Methoden für automatische Sequenzanalysen in der historischen Linguistik gegeben werden. Vorgestellt werden dabei grundlegende Aspekte von Sequenzanalysen in der historischen Linguistik, traditionelle und computergestützte Methoden der Alinierung, die bisher geleisteten Vorarbeiten, sowie Ziele und Pläne für das weitere Vorgehen.

1 Grundlegende Aspekte

1.1 Deterministische und heuristische Ansätze

In Bezug auf automatische Sequenzanalysen in der historischen Linguistik kann grundlegen zwischen deterministischen und heuristischen Ansätzen unterschieden werden. Während deterministische Ansätze versuchen, auf Grundlage eines eindeutigen Regelsystems Sprachwandelprozesse zu modellieren (vgl. bspw. das Programm zur Ableitung des Polnischen aus dem Urslawischen in Kondrak 2002), besteht das Ziel heuristischer Ansätze im Auffinden dieser Prozesse, die in den deterministischen Ansätzen als gegeben angenommen werden.

1.2 Alinierung als grundlegende Methode der Sequenzanalyse

Da es sich bei Sequenzen um geordnete Symbolketten handelt, deren einzelne Symbole Distinktivität erst aufgrund ihrer spezifischen Anordnung erlangen, stellt die Alinierung die grundlegende Methode für die Sequenzanalyse dar. Im Gegensatz zu Mengenvergleichen, die auf dem Vergleich ungeordneter Kollektionen von Elementen beruhen, ist für Sequenzanalysen die Anordnung der Elemente von entscheidender Bedeutung. Daraus ergibt sich, dass für Sequenzanalysen grundlegend unterschiedliche Verfahren angewendet werden müssen, als für Mengenvergleiche.

Language	Word	Meaning
Mandarin	$ma_{55}ma_{3}$	“mother”
German	mama	“mother”
Russian	tak	“in this way”
German	$t^{h}a{:}k$	“day”
Russian	mif	“myth”
German	mi:f	“stale air”

Language	Word	Meaning
German	$ts^{h}a{:}n$	“tooth”
English	tu:θ	“tooth”
Ukrainian	jasni	“gums”
Russian	dɨsna	“gums”
English	ma:lboʁo	“Marlboro”
Mandarin	$wan_{51}paw_{21}lu_{51}$	“Marlboro”

Tabelle 1: Synchrone und diachrone Ähnlichkeit

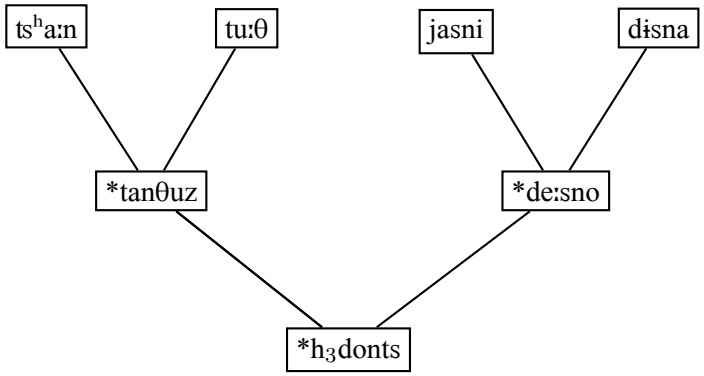

Abbildung 1: Entwicklunsszenario als Erklärungsansatz für funktionale Ähnlichkeit

2 Ähnlichkeit

2.1 Synchrone und diachrone Ähnlichkeit

Für Zwecke der automatischen Sequenzanalyse in der historischen Linguistik muss strikt zwischen einer synchronen und einer diachronen Auffassung von Ähnlichkeit unterschieden werden: Während synchrone Ähnlichkeit Laute unterschiedlicher Sprachen als ähnlich einstuft, wenn diese Gemeinsamkeiten in Bezug auf Artikulation oder Perzeption aufweisen, beruht diachrone Ähnlichkeit lediglich auf einer funktionalen Entsprechung der Laute. Die funktionale Entsprechung ist hierbei zunächst vollkommen unabhängig von einer wie auch immer definierten synchronen Ähnlichkeit lautlicher Segmente. Die linke Spalte von Tabelle 1 gibt drei Beispiele für Wörter, die aus einer synchronen Perspektive ähnlich sind. Diese Wörter sind allesamt nicht miteinander verwandt, d. h. sie gehen nicht auf einen gemeinsamen Vorgänger zurück. Demgegenüber zeigt die rechte Spalte von Tabelle 1 Wörter, die auf eine gemeinsame Vorgängerform zurückgehen: Hier kann nicht mehr von einer synchronen Ähnlichkeit der Segmente die Rede sein, vielmehr muss ein auf Lautwandelprozessen basierendes Entwicklungsszenario angenommen werden, dass die funktionalen Entsprechungen der lautlichen Segmente begründet (vgl. Abbildung 1).

2.2 Lautwandel und Lautkorrespondenz

Die funktionale Ähnlichkeit zweier lautlicher Segmente wird in der historischen Linguistik im Rahmen der Suche nach regulären Lautkorrespondenzen festgestellt: Wenn sich eine in ihrem Umfang meist nicht definierte Menge von Wörtern verschiedener Sprachen finden lässt, in denen lautliche Segmente miteinander korrespondieren, so werden diese Segmente als funktional ähnlich eingestuft und es wird davon ausgegangen, dass diese funktionale Ähnlichkeit historisch erklärt werden kann, sei es durch die Annahme einer gemeinsamen Ursprache, aus der die untersuchten Sprachen entstanden sind, oder durch von Sprachkontakt (vgl. hierzu auch Lass 1997, 123f). Obwohl in dieser Konzeption funktioneller Entsprechung die Ebene der synchronen Ähnlichkeit irrelevant ist (was oft in der linguistischen Literatur betont wird), so geht man in der historischen Linguistik schon seit längerem davon aus, dass Lautwandelprozesse sich in einem bestimmten Rahmen vollziehen, der eine phonetisch-phonologische Komponente hat. Der strikten Annahme, dass jeder Laut mit jedem korrespondieren könne, unabhängig von seiner tatsächlichen Artikulation, kann somit durch eine stochastische Komponente abgeschwächt werden, welche besagt, dass bestimmte Arten von Lautkorrespondenzen häufiger zu verzeichnen sind, als andere. Diese weniger starke Formulierung des Prinzips der funktionalen Ähnlichkeit lässt sich für heuristische Ansätze der automatischen Sequenzanalyse in der historischen Linguistik ausnutzen.

2.3 Die Direktionalität von Lautwandelprozessen

Ein weiterer wichtiger Aspekt von Lautwandelprozessen in der historischen Linguistik ist deren Direktionalität. Eine Vielzahl von Lautwandelprozessen ist in ihrer grundlegenden Struktur gerichtet: Während velare Laute relativ häufig palatalisiert werden, ist der umgekehrte Prozess kaum zu verzeichnen. Dies trifft nicht auf alle Arten von Lautwandelprozessen zu, gerichtete Lautwandelprozesse treten jedoch relativ häufig auf und müssen bei der Erstellung von Methoden zur automatischen Sequenzanalyse in jedem Fall berücksichtigt werden.

2.4 Lautklassen

Ein früher Ansatz zur stochastischen Behandlung von Lautwandelprozessen in der historischen Linguistik findet sich bei Dolgopolsky (1986). Auf Grundlage einer Sichtung von Lautkorrespondenzen in einem Datenset von etwa 400 Sprachen stellte dieser 10 Lautklassen auf, unter welchen gängige Laute in den Sprachen der Welt subsumiert werden (vgl. Tabelle 2). Die Grundannahme Dolgopolskys war dabei, dass es möglich sei

> [...] to divide sounds into such groups, that changes within the boundary of the groups are more probable than transitions from one group into another. (Burlak & Starostin 2005, 272)[1]

Lautwandel innerhalb der Klassen wird also als wahrscheinlicher angenommen, als zwischen den Klassen. Das Lautklassenkonzept stellt somit einen Versuch dar, die strikte Auffassung funktionaler Ähnlichkeit in der historischen Linguistik durch eine stochastische Komponente zu erweitern.

Dolgopolsky selbst formulierte seine Lautklassen als absolut, Übergangswahrscheinlichkeiten zwischen den Klassen wurden nicht angenommen, jedoch lässt sich sein Ansatz in dieser Hinsicht relativ

[1] Meine Übersetzung, Originaltext: «[...] выделить такие группы звуков, что изменения в пределах группы более вероятны, чем переводы из одной группы в другую.»

No.	Class	Description	Example
1	P	labial obstruents	p,b,f
2	T	dental obstruents	d,t,θ,ð
3	S	sibilants	s,z,ʃ,ʒ
4	K	velar obstruents, dental and alveolar affricates	k,g,ʦ,ʧ
5	M	labial nasal	m
6	N	remaining nasals	n,ɲ,ŋ
7	R	liquids	r,l
8	W	voiced labial fricative and initial rounded vowels	v,u
9	J	palatal approximant	j
10	ø	laryngeals and initial velar nasal	h,ɦ,ŋ

Tabelle 2: Lautklassenschema von Dolgopolsky

leicht verfeinern, indem man die Anzahl der Lautklassen erhöht und anstelle der absoluten Identitätsannahme Übergangswahrscheinlichkeiten für Klassenübergänge ansetzt. Der Vorteil von lautklassenbasierten Ansätzen zur automatischen Sequenzanalyse liegt neben deren Flexibilität insbesondere auch darin, dass Lautklassenalphabete relativ klein sind und somit einfach mit Hilfe biologischer Softwarepakete analysiert werden können. Tabelle 3 zeigt ein erweitertes Lautklassenschema, das für die derzeitigen Berechnungen vorläufig angesetzt wird.

3 Alinierung

3.1 Traditionelle Alinierungsmethoden

Im Rahmen der traditionellen historischen Linguistik wurde explizit nie von Alinierung gesprochen, jedoch liegt der Postulierung regulärer Lautkorrespondenzen zwangsläufig eine Alinierung zugrunde (vgl. Anttila 1972), da diese nur durch eine Alinierung von Sequenzen ermittelt werden können. Jedoch beschränkt sich die traditionelle Analyse auf einen weitestgehend qualitativen Vergleich. Es bleibt der Intuition des Forschers überlassen, welche Segmente er welchen gegenüberstellt. Dass jedoch Elemente gegenübergestellt werden, steht hierbei außer Frage. Werden bspw. im Rahmen der historischen Linguistik die beiden Sequenzen gr. [θiɣatɛra] und engl. [dɔtɚ] “Tochter” verglichen, so ist ein Feststellen der Lautkorrespondenzen gr. [θ] : engl. [d] bzw. gr. [t] : engl. [t] nur im Rahmen einer Alinierung zu realisieren (vgl. Abbildung 2).

3.2 Der ahistorische Charakter von Alinierungen

Alinierungen haben grundlegend einen ahistorischen Charakter. Sei enthalten keine historische Aussage darüber, wie die Entsprechungen historisch zu erklären sind. Eine historische Interpretation erlangen Alinierungen erst im Rahmen der linguistischen Rekonstruktion (vgl. Fox 1995), wenn eine explizite Vorgängerform der Sequenzen postuliert wird, aus deren Struktur sich ein Entwicklungsschema für die alinierten Sequenzen ableiten lässt. Das Konzept des Rekonstrukts unterscheidet sich in dieser Hinsicht

No.	Klasse	Beschreibung	Bsp.
1	P	labiale Plosive	p, b
2	F	labiale Frikative	f, β
3	S	alveolare, retroflexe und postalveolare Frikative	s, z, ʃ, ʒ
4	K	velare und uvulare Plosive	k, g
5	G	velare und uvulare Frikative	x, ɣ
6	C	Affrikaten	ʦ, ʧ
7	M	labialer Nasal	m
8	R	Trills, Taps, Flaps	r
9	L	Laterale Approximanten	l
10	N	Nasale	n, ŋ
11	W	labialer Approximant, stimmhafter labialer Frikativ	w,v
12	J	palataler Approximant	j
13	H	Laryngale	h, ʔ
14	T	dentale und alveolare Plosive	t, d
15	D	dentale und alveolare Frikative	θ, ð
16	A	offene, ungerundete Vokale	a, ɑ
17	E	mittlere, ungerundete Vokale	e, ɛ
18	I	geschlossene, ungerundete Vokale	i, ɪ
19	O	gerundete, offene Vokale	o,ɔ
20	U	gerundete, geschlossene Vokale	u, ʊ

Tabelle 3: Derzeitiges Lautklassenschema

grundlegend von dem in der Biologie gebräuchlichen Konzept der “Consensus-Sequenz”: Während die Consensus-Sequenz eine Metadarstellung einer multiplen Alinierung hinsichtlich der am häufigsten auftretenden Segmente in bestimmten Positionen darstellt, ist für das Rekonstrukt in der historischen Linguistik die Häufigkeit der Segmente in bestimmten Positionen nicht entscheidend, sondern die historische Erklärungskraft des Rekonstrukts, also die Frage, durch welche Protoform sich ein historisches Entwicklungsszenario am besten begründen lässt. Da die Rekonstruktion hierbei auf einer Rückführung von Lautwandelprozessen beruht, ist die Frage der Gerichtetheit von Lautwandelprozessen entscheidend für dieses Verfahren.

3.3 Grundlegende Methoden der automatischen Alinierung

Beginnend mit Levenshteins Formulierung der Edit-Distanz als Distanzmaß für paarweise Sequenzvergleiche (vgl. Levenshtein 1966), wurden ab den Siebzigern von verschiedenen Forschern unabhängig voneinander algorithmische Verfahren zur Ermittlung von Sequenzdistanzen und zur automatischen Alinierung vorgeschlagen (vgl. Needleman & Wunsch 1970, Wagner & Fischer 1974), die in ihren Grundlagen allesamt auf dem dynamischen Programmieralgorithmus beruhen, der auch die Grundlage für die gängigen Verfahren zur automatischen Sequenzanalyse in der Biologie darstellt. Aufbauend auf diesem Algorithmus werden insbesondere in der Evolutionsbiologie sowohl paarweise als auch multiple Alinie-

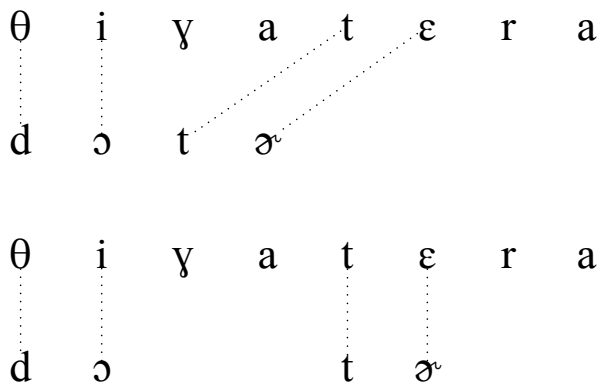

Abbildung 2: Alinierung als Grundlage der Ermittlung von Lautkorrespondenzen

rungen automatisch ermittelt. In der historischen Linguistik wird der Algorithmus abgesehen von einigen wenigen Ansätzen zur Sequenzalinierung (vgl. Kondrak 2002) meist für die automatische Ermittlung von Distanzen zwischen Sprachen und Dialekten verwendet (vgl. Heeringa *et al.* 2006, Holman *et al.* 2008b, Downey *et al.* 2008), der Alinierungsaspekt ist hierbei zweitrangig.

3.3.1 Paarweise Alinierung

Die grundlegende Idee des dynamischen Programmieralgorithmus zur automatischen Sequenzalinierung besteht im Erstellen einer Matrix, in der alle Segmente zweier Sequenzen einander gegenübergestellt werden. Der Algorithmus sucht nun schrittweise den kürzesten Weg von der rechten oberen Ecke zur linken unteren Ecke der Matrix, wobei mit Hilfe einer speziellen Scoring-Funktion, welche das Gegenüberstellen, das Auslassen oder das Hinzufügen von Segmenten bewertet, kumulativ die Gesamtkosten der Alinierung aufgerechnet werden. Der Weg mit dem geringsten (im Falle von Distanzberechnungen) oder höchsten (im Falle von Ähnlichkeitsberechnungen) Wert stellt dabei die optimale Alinierung beider Sequenzen dar. Alternativ lässt sich dabei auch im Rahmen einer lokalen Alinierung diejenige Subsequenz zwischen beiden Sequenzen ermitteln, welche die höchste Ähnlichkeit aufweist (vgl. Smith & Waterman 1981).

Abbildung 3 zeigt am Beispiel der Wörter engl. “heart” und dt. “herz”, wie der dynamische Programmieralgorithmus zunächst eine Matrix erstellt, in der alle Segmente (Gaps eingeschlossen) einander gegenübergestellt werden (linke Abbildung) und dann durch eine Scoring-Funktion (die Levenshtein-Distanz im vorliegenden Beispiel) kumulativ der Pfad durch die Matrix mit den geringsten Kosten berechnet wird. Die Sequenzdistanz (2 in diesem Fall) findet sich dabei in der rechten unteren Zelle der Matrix wieder.

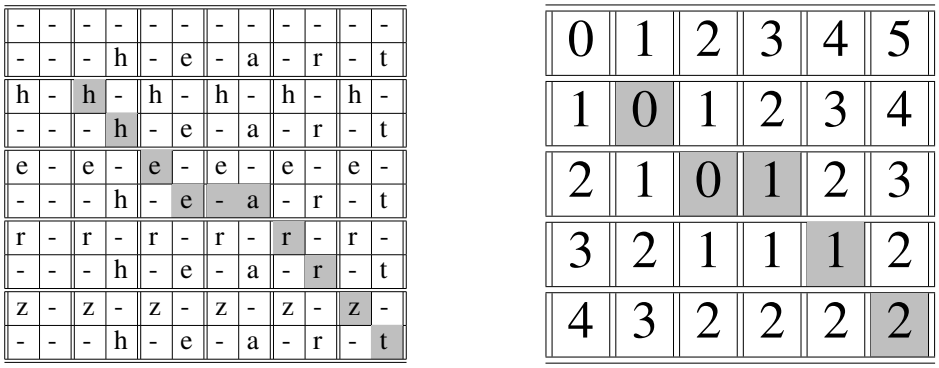

Abbildung 3: Matrix für die Alinierung von engl. “heart” und dt. “herz”

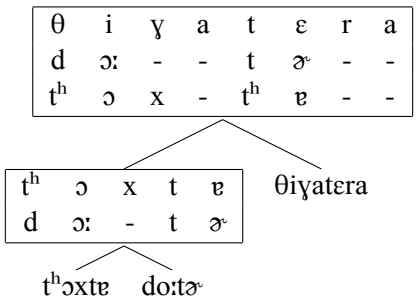

Abbildung 4: Progressive Alinierung anhand eines Leitbaums

3.3.2 Multiple Alinierung

Da der Rechenaufwand für eine Berechnung multipler Alinierungen mit Hilfe des dynamischen Programmieralgorithmus potentiell ansteigt, und somit ab einer Alinierung von mehr als drei Sequenzen nicht mehr praktikabel ist, werden in der Biologie meist spezielle Heuristiken zur Ermittlung multipler Alinierungen verwendet. Im Rahmen progressiver Methoden zur multiplen Alinierung, die meist eine Variante des Feng-Doolittle-Algorithmus darstellen (Feng & Doolittle 1987), wird dabei auf Grundlage einer Berechnung paarweiser Sequenzdistanzen mit Hilfe von Clusteralgorithmen zunächst ein Leitbaum (*guide-tree*) erzeugt, anhand dessen die Sequenzen dann sukzessive zu einer multiplen Alinierung vereinigt werden. Es gibt jedoch keine Garantie dafür, dass die optimale multiple Alinierung durch dieses Verfahren gefunden werden kann. Die Güte einer derartigen Alinierung hängt starkt von der Struktur der Sequenzen ab, die miteinander aliniert werden. Auch kann die Konstruktion des Leitbaums erheblichen Einfluss auf die Alinierung der Sequenzen nehmen.

Abbildung 4 zeigt anhand der Sequenzen gr. θιγατέρα [θiɣatɛra], engl. *daughter* [dɔːtɐ̯] und dt. *Tochter* [tʰɔxtʰɐ] “Tochter”, wie eine multiple Alinierung anhand eines Leitbaums erfolgt. Dabei gibt der Leitbaum zunächst eine Alinierung des englischen und des deutschen Wortes vor, welcher dann das griechische Wort

hinzugefügt wird.

3.3.3 Die Scoring-Funktion

Abgesehen von Erweiterungen, die in den dynamischen Programmieralgorithmus selbst eingreifen, um bestimmten Phänomenen wie Metathese (vgl. den Damerau-Levenshtein-Algorithmus, dessen theoretische Grundlage in Damerau 1964 formuliert wurde), konsekutiven Gaps (vgl. Gotoh 1982) oder Kontraktionen und Expansionen (vgl. Oommen 1995) in der Berechnung gerecht zu werden, kommt der Scoring-Funktion eine ungleich wichtige Rolle zu, da durch diese das Gegenüberstellen von Segmenten geregelt wird. Grundlegend liegt der Scoring-Funktione eine sehr einfache Struktur zugrunde: sie liefert für zwei Segmente ein Distanz- oder ein Ähnlichkeitsmaß zurück. Eine der einfachsten Scoring-Funktionen stellt dabei die Levenshtein-Distanz dar, welche für identische Segmente die Distanz 0 liefert und für nichtidentische Segmente (Gaps eingeschlossen) die Distanz 1. Obwohl für die allgemeine Berechnung von Sprachdistanzen auf dieser Basis bereits recht gute Ergebnisse erzielt werden können (vgl. Serva & Petroni 2008, Holman *et al.* 2008a), bedarf es für Ansätze, bei denen die Alinierung und nicht eine allgemeine lexikalische Distanz zwischen Sprachen im Vordergrund steht, verfeinerter Scoring-Funktionen, um zufriedenstellende Ergebnisse zu erzielen. Kompliziertere Scoring-Funktionen gehen von Substitutionsmatrizzen aus, die für jede Gegenüberstellung zweier Segmente je einen individuellen Wert liefern. Wichtig ist in diesem Zusammenhang, dass dem funktionalen Ähnlichkeitskonzept der historischen Linguistik in seiner stochastischen Erweiterung Rechnung getragen wird. Eine lediglich auf synchroner Lautähnlichkeit beruhende Scoring-Funktion ist für historische Ansätze nicht angebracht.

3.4 Erweiterte Ansätze

3.4.1 Profilbasierte Alinierung

Während im klassischen Feng-Doolittle-Algorithmus die Alinierung von multiplen mit multiplen Alinierungen auf Grundlage der paarweisen Alinierung der Sequenzen mit der größten Ähnlichkeit vorgenommen wird, die somit als stellvertretend für eine multiple Alinierung genommen werden, gibt es ferner die Möglichkeit, eine multiple Alinierung durch ein sogenanntes "Profil" wiederzugeben. Ein Profil beinhaltet die relative Häufigkeit des Auftretens aller Segmente der alinierten Sequenzen in einer bestimmten Position der multiplen Alinierung. Da ein Profil eine multiple Alinierung durch eine Sequenz von Vektoren darstellt, lassen sich somit Profile paarweise alinieren. Es herrscht jedoch in der Biologie kein einheitlicher Ansatz für die Beschaffenheit der Scoring-Funktion vor, welche die entsprechenden Vektoren der Profile miteinander vergleicht (vgl. Edgar & Sjolander 2004).

Tabelle 4 gibt ein Beispiel für die Erstellung eines Profils aus der multiplen Alinierung der drei Sequenzen gr. θυγατέρα [θiɣatɛra], engl. *daughter* [dɔ:tɚ] und dt. *Tochter* [tʰɔxtɐ] "Tochter". Die IPA-kodierten Laute werden dabei zunächst in Lautklassen (vgl. Tabelle 3) überführt. Daraufhin wird für jede Position in der multiplen Alinierung die relative Häufigkeit des Auftretens jeder Lautklasse berechnet.Das so entstandene Profil stellt eine Sequenz aus Vektoren dar, welche Informationen bezüglich der relativen Häufigkeit der Einzelsegmente, die in der multiplen Alinierung vorkommen, enthält.

Dass die profilbasierte Alinierung im Gegensatz zur traditionellen Alinierung nach dem Feng-Doolittle-Algorithmus Vorteile aufweist, lässt sich leicht am Beispiel der multiplen Alinierung der Sequenzen tsch. *člověk* [ʧlovʲɛk], bulg. човек [ʧovɛk], russ. человек [ʧʲɪlɐvʲɛk] und poln. *człowiek* [ʧwɔvʲɛk] "Mensch" aufzeigen. Während, wie Abbildung 5 zeigt, die Profilalinierung poln. [w] dem Laut [l] in den übrigen Sprachen richtig gegenüberstellt, aliniert die reine Leitbaumalinierung poln. [w] mit russ. [ɪ] in zweiter

	Multiple Alinierung: Ausgabeformat							
	θ	i	ɣ	a	t	ɛ	r	a
	d	ɔː	-	-	t	ɤ̞	-	-
	tʰ	ɔ	x	-	tʰ	ɐ	-	-
	Multiple Alinierung: Lautklassenformat							
	D	I	G	A	T	E	R	A
	T	O	-	-	T	E	-	-
	T	O	G	-	T	A	-	-
	Multiple Alinierung: Profil							
A				.33		.33		.33
E						.66		
O		.66						
I		.33						
D	.33							
T	.66				1.0			
G			.66					
R							.33	
-			.33	.66			.66	.66

Tabelle 4: Profil für eine multiple Alinierung

Position. Der Grund hierfür liegt darin, dass für die Vereinigung multipler Alinierungen lediglich die beiden ähnlichsten Sequenzen stellvertretend für die gesamte Alinierung miteinander paarweise aliniert werden. Dabei können wichtige Informationen aus den Gesamtalinierungen verlorengehen. So bestraft die derzeitige Substitutionsmatrix das Gegenüberstellen von Vokalen und Konsonanten mit hohen negativen Kosten (-10). Da russ. [ɪ] in der paarweisen Alinierung der beiden Stellvertretersequenzen jedoch nicht auftaucht, werden lediglich die Kosten für ein Gegenüberstellen von Gaps mit Gaps angesetzt, welche traditionell mit 0 angesetzt werden.

3.4.2 Trigram-basierte Alinierung

Während in biologischen Ansätzen zur Sequenzalinierung weitestgehend von einer Unabhängigkeit der Segmente von den sie umgebenden Elementen ausgegangen wird, ist in Bezug auf den Lautwandel die lautliche Umgebung, in der sich bestimmte Lautwandelprozesse vollziehen, von großer Bedeutung. Eine einfache Möglichkeit, um der Abhängigkeit lautlicher Segmente von Vorgänger- und Nachfolgersegmenten gerecht zu werden, stellt Überführung der Sequenzen in Trigramme dar. Hierbei wird jedes Segment jeweils in einem Tuple aus Vorgängersegment, Segment und Nachfolgersegment dargestellt. Für die Alinierung derartiger Sequenzen muss die Scoring-Funktion dahingehend modifiziert werden, dass sie aus allen drei Elementen des Tuples ein Distanz- oder Ähnlichkeitsmaß errechnet. Der Vorteil eines solchen Ansatzes besteht in der Einbeziehung von kontextuellen Informationen, die bei der Identifizierung diachroner Ähnlichkeiten in der historischen Linguistik spätestens seit Karl Verners Erweiterung des Grimmschen Gesetzes zum Standard geworden ist. Geklärt werden muss bei trigrambasierten Ansätzen allerdings die Frage, wie genau der Wert für die Gegenüberstellung von Segmenten berechnet werden

Profilbasierte Alinierung							
Czech (1)	ʧ	-	l	o	vʲ	ɛ	k
Bulgarian (2)	ʧ	-	-	o	v	ɛ	k
Russian (3)	ʧʲ	ɪ	l	ɐ	vʲ	ɛ	k
Polish (4)	ʧ	-	w	ɔ	vʲ	ɛ	k
Leitbaumbasierte Alinierung							
Czech (1)	ʧ	-	l	o	vʲ	ɛ	k
Bulgarian (2)	ʧ	-	-	o	v	ɛ	k
Russian (3)	ʧʲ	ɪ	l	ɐ	vʲ	ɛ	k
Polish (4)	ʧ	w	-	ɔ	vʲ	ɛ	k

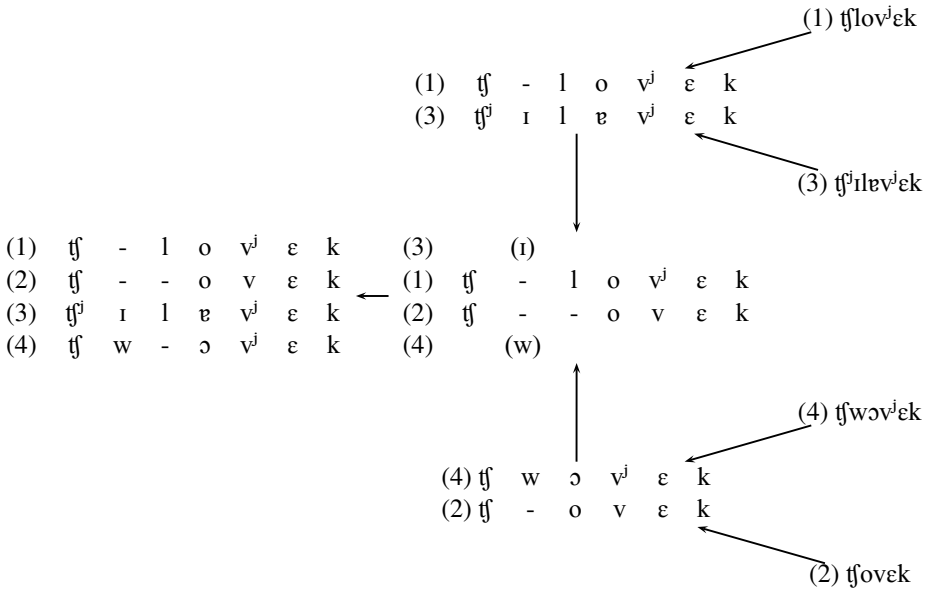

Abbildung 5: Probleme nicht profilbasierter Leitbaumalinierung

soll. Eine zu starke Gewichtung von Vorgänger- und Nachfolgersegment kann die Ergebnisse leicht verfälschen, genauso wie eine zu niedrige Gewichtung in vielen Fällen überhaupt keinen Einfluss auf die Alinierung haben wird.

Tabelle 5 gibt ein Beispiel für eine auf Trigrambasis erstellte multiple Alinierung der Sequenzen gr. θιγατέρα [θiɣatɛra], engl. *daughter* [dɔːtɚ] und dt. *Tochter* [tʰɔxtɐ] "Tochter". Die Ausgansstrings wurden dabei zunächst in die derzeit verwendeten Lautklassen (vgl. Tabelle 3), und daraufhin in Trigramme umgewandelt, wobei als Vorgängerumgebung für das erste und letzte Segment der Trigramme jeweils ein Gap-Zeichen angesetzt wurde (diese Notation kann im weiteren Verlauf verfeinert und eine explizite Notation für Wortanfang und Wortende definiert werden, an welche die Scoring-Funktion angepasst werden

θ	i	ɣ	a	t	ɛ	r	a
d	ɔː	-	-	t	ɐ̯	-	-
t^h	ɔ	x	-	t	ɐ	-	-
D	I	G	A	T	E	R	A
T	O	-	-	T	E	-	-
T	O	G	-	T	A	-	-
(-, D, I)	(D, I, G)	(I, G, A)	(G, A, T)	(A, T, E)	(T, E, R)	(E, R, A)	(R, A, -)
(-, T, O)	(T, O, T)	-	-	(O, T, E)	(T, E, -)	-	-
(-, T, O)	(T, O, G)	(O, G, T)	-	(G, T, A)	(T, A, -)	-	-

Tabelle 5: Trigrambasierte multiple Alinierung

kann). Als Scoring-Funktion wurde eine einfache Funktion angesetzt, welche die Bewertungen für die drei Elemente der jeweiligen Tuples aufsummiert, die Bewertung für Vorgänger- und Nachfolgersegment jedoch lediglich mit zehn Prozent gewichtet.

4 Bisherige Vorarbeiten

Um die oben genannten Vorüberlegungen umzusetzen und zu testen, wurde ein Software-Paket mit speziellen Bibliotheken in der Skriptsprache Python erstellt, das verschiedene bereits postulierte und neue Methoden zur Sequenzanalyse zur Verfügung stellt. Grundlegendes Eingabeformat für die alle Methoden in den Bibliotheken sind IPA-kodierte Strings. IPA wurde gewählt, um ein möglichst einheitliches Format der phonetischen Kodierung zu ermöglichen. Um die Vergleichbarkeit zu den von anderen Autoren postulierten Algorithmen, die im Programm implementiert sind, zu gewährleisten, wurden Ersetzungsfunktionen geschrieben, die das IPA-Format in das von den jeweiligen Autoren verwendete Format überführen. Das Programm erlaubt gegenwärtig die Durchführung paarweiser und multipler Alinierungen. Multiple Alinierungen können ferner auf Basis des traditionellen Feng-Doolittle-Algorithmus durchgeführt werden, wie auch auf Profilbasis. Ferner sind trigrambasierte paarweise und multiple Alinierungen möglich.

4.1 Scoring-Funktionen für Lautklassen, Sequenzprofile und Trigramme

In der Biologie basieren die Scoring-Funktionen für Sequenzalinierungen meist auf Substitutionsmatrizzen, die auf empirischer Basis erstellt wurden und Aussagen über die grundlegende Wahrscheinlichkeit enthalten, dass zwei Segmente (Proteine, Aminosäuren) in einer (korrekten) Alinierung einander gegenübergestellt werden (Rauhut 2001, 42-49). Da die traditionelle historische Linguistik bis heute weitgehend ein qualitatives Vorgehen beim Auffinden von Lautkorrespondenzen aufrechterhält, in dem Lautkorrespondenzen eher absolut postuliert denn tatsächlich ausgezählt werden und das Auffinden neuer Lautkorrespondenzen weitestgehend der Intuition des jeweiligen Forschers überlassen wird (Schwink 1994, 29), ist es zum jetzigen Zeitpunkt nicht möglich, ein rein empirisch basiertes stochastisches Modell von Lautübergängen aufzustellen. Für die ersten Untersuchungen muss daher ein weitgehend auf eigener Kenntnis von Lautwandelprozessen beruhendes vorläufiges Modell entwickelt werden, das in einem weiteren

Schritt an Datensets größeren Umfangs getestet und verfeinert werden muss.

4.1.1 Scoring-Funktionen für Lautklassen

Bei den Scoring-Funktionen für Lautklassen sind folgende Vorüberlegungen entscheidend: Zunächst muss das auf lediglich zehn Lautklassen basierende Schema von Dolgopolsky (1986) für die Alinierung verfeinert werden, um komplexere Strukturen entdecken zu können. Während Dolgopolskys ursprüngliche Idee keine strikte Alinierung, sondern lediglich den Vergleich der ersten beiden konsonantischen Segmente in Sequenzen vorsah, ermöglichen die Alinierungsalgorithmen in Verbindung mit den Scoring-Funktionen eine genauere Heuristik für ähnliche Strukturen in Sequenzen. In Bezug auf den prinzipiell ahistorischen Charakter von Alinierungen, der im Gegensatz zum vielfach direktionalen Charakter von Lautwandelprozessen steht, ist es wichtig, eine Scoring-Funktion zu erstellen, die in ihren Grundzügen nicht metrisch ist. Während sich in einer metrischen Scoring-Funktion die Distanz von zwei Segmenten A und B zu einem Segment C abhängig von der Distanz zwischen A und B ist (so kann aufgrund der Dreiecksungleichheit bspw. die Distanz von A zu C nicht größer sein als die Summe der Distanz von A und B und B und C), muss dies in einer nicht-metrischen Scoring-Funktion nicht zwangsläufig gegeben sein. Diese nicht-metrische Struktur ist bei gerichteten Sprachwandelprozessen von großer Bedeutung: So ist der Wandel von Velaren ebenso wie der Wandel von Dentalen zu Affrikaten ein häufig auftretendes Phänomen. Wenn diese Prozesse nicht gerichtet wären, wäre für einen Wandel von Dentalen zu Velaren somit ebenfalls als relativ häufig auftretender Lautwandelprozess anzunehmen. Dies ist aufgrund der Gerichtetheit der beiden Prozesse nicht der Fall. Für eine Scoring-Funktion, die dieser Tatsache Rechnung trägt, muss demnach die Distanz zwischen Dentalen und Velaren viel höher angesetzt werden, als die von Affrikaten zu Velaren und Dentalen.

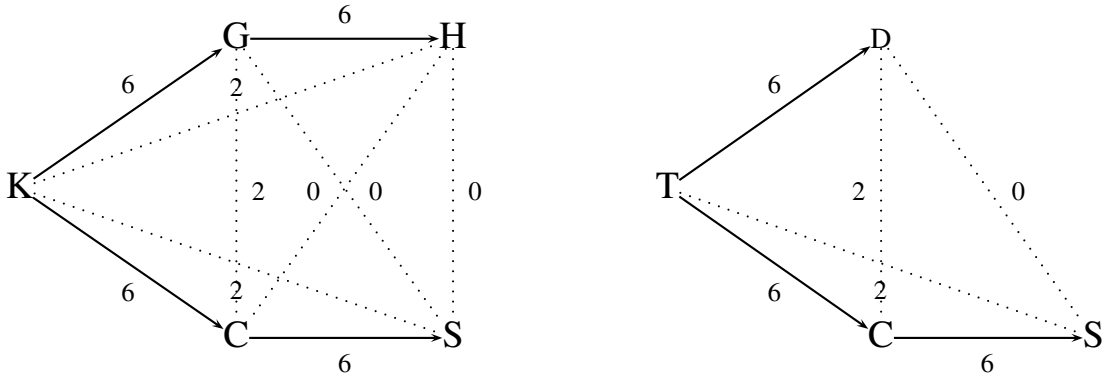

Abbildung 6: Gerichtete Lautwandelprozesse zwischen Lautklassen

Zur Erstellung einer derartigen Scoring-Funktion, die dem gerichteten Charakter verschiedener häufiger Lautwandelprozesse Rechnung trägt, bietet es sich an, die Lautwandelprozesse in voneinander unabhängigen Bäumen abzubilden, wobei die Wurzeln jeweils einen möglichen Ausgangspunkt eines Prozesses darstellt. Für das erste Testen von lautklassenbasierten Ansätzen werden in der jetzigen Form des Programms, wie bereits erwähnt, insgesamt 20 Lautklassen angesetzt (vgl. Tabelle 3). Will man nun erste rudimentäre Übergangswahrscheinlichkeiten für die Lautklassen *T* und *K* darstellen, so lassen sich diese in zwei getrennten Bäumen darstellen. Wenn wir von einer Scoring-Funktion ausgehen, die auf Ähnlichkeit beruht, wie dies in der Biologie allgemein üblich ist, und die Ähnlichkeit für Identität von Konsonanten, die der selben Klasse angehören auf 10 festlegen, so können in einem weiteren Schritt, auf Basis eines Lautwandelschemas, wie in Abbildung 6 dargestellt, Ähnlichkeiten für weitere Lautklassen

definiert werden, in welche diese übergehen können. Da der Übergang von *K* nach *S* dabei allgemein über den Zwischenschrit *K* > *C* erfolgt, wird die Ähnlichkeit zwischen den beiden Segmenten entsprechend niedriger (mit 2) angesetzt. Die Unabhängigkeit beider Bäume für *T* und *K* garantiert dabei, dass die beiden Ursprungssegmente nicht selbst miteinander in Beziehung gesetzt werden (die Ähnlichkeit wird hier grundlegend mit 0 angesetzt). Die Ähnlichkeitsbeziehungen, die für die einzelnen Lautklassen postuliert werden, müssen in diesem Zusammenhang nicht zwangsläufig in einem Baumschema dargestellt werden. In Fällen, in denen eine Direktionalität von Lautwandelprozessen nicht mit hoher Wahrscheinlichkeit begründet werden kann, lassen sich auch netzartige Strukturen darstellen. Wichtig ist lediglich die prinzipielle Unabhängigkeit der Entwicklungsbäume voneinander, d.h. die Tatsache, dass *K* zu *S* in einer gewissen Ähnlichkeitsbeziehung steht, begründet nicht die Postulierung einer Ähnlichkeitsbeziehung für *K* und *T*, nur weil *T* ebenfalls in einer gewissen Ähnlichkeitsbeziehung zu *S* steht.

Eine erste vorläufige Scoring-Funktion für die 20 Lautklassen wurde bereits entwickelt und wird gegenwärtig getestet. Für das erweiterte Training der Scoring-Funktion müssen allerdings Daten in größerem Umfang als bisher einbezogen werden, da nur eine große Datenbasis, die möglichst Sprachen aus verschiedenen Sprachfamilien einbeziehen sollte, die Qualität einer derartigen Heuristik gewährleisten kann.

4.1.2 Scoring-Funktionen für Sequenzprofile

Die derzeitige Scoring-Funktion für die Profil-Profil-Alinierung basiert auf dem *sum of pairs score* (vgl. Durbin 2002, 146f), der in einigen biologischen Programmen zur multiplen Sequenzalinierung verwendet wird (beispielsweise in CLUSTAL W, vgl. Thompson *et al.* 1994, 4675). Hierbei werden die Werte, welche die Substitutionsmatrix für alle Kombinationen der Segmente zweier multipler Alinierungen zurückgibt, aufsummiert und anschließend deren Durchschnitt berechnet. Für zwei Positionen P_x und P_y der Länge m und n in zwei multiplen Alinierungen M_x und M_y, welche die Segmente $x_1, x_2, ..., x_m$ und $y_1, y_2, ..., y_n$ enthalten ergibt sich für eine Scoring-Funktion σ als Formel zur Berechnung der Bewertung somit die folgende Formel:

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\sigma(x_i, y_j)}{mn} \quad (1)$$

Angenommen, wir haben zwei Positionen in zwei multiplen Alinierungen, von denen die erste die drei Segmente *T*, *V* und *C* enthält und die zweite die Elemente *T* und *S*, so berechnet sich die Bewertung der Positionen, wenn man die derzeit in dem Programm verwendete Substitutionsmatrix für paarweise Alinierung ansetzt, wie folgt:

$$\frac{\sigma(T,T)+\sigma(T,S)+\sigma(V,T)+\sigma(V,S)+\sigma(C,T)+\sigma(C,S)}{2*3} = \frac{10+2+6+0+6+6}{6} = 5 \quad (2)$$

4.2 Lautklassenbasierte paarweise Alinierung

Lautklassenbasierte Alinierung wird derzeit mit Hilfe von Bibliotheken, die für biologische Analysen in Python erstellt wurden (BioPython, vgl. Cock *et al.* 2009 und PyCogent, vgl. Knight *et al.* 2007), realisiert, die um die oben erwähnten Scoring-Funktionen ergänzt wurden, um den linguistischen Ansprüchen gerecht zu werden. Der Rückgriff auf Implementierungen, die für biologische Zwecke erstellt wurden, ist aus zwei Gründen von Vorteil: Zunächst erspart sie aufwendige Programmierarbeit und Einarbeitung in Algorithmen, die in der jetzigen Form professionell implementiert wurden, des Weiteren sind die Algorithmen der erwähnten Bibliotheken in C++ realisiert und reinen auf Python basierenden Algorithmen

an Geschwindigkeit überlegen. Ferner weisen die Pakete eine hohe Flexibilität auf und lassen sich somit einfach an linguistische Zwecke anpassen. Abbildung 7 zeigt das grundlegende Vorgehen im Rahmen der

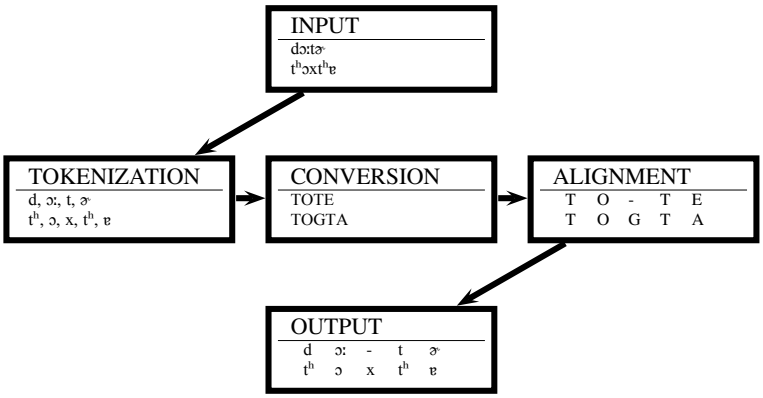

Abbildung 7: Grundsätzliche Arbeitsweise der lautklassenbasierten Alinierung

lautklassenbasierten Alinierung am Beispiel der Alinierung der Sequenzen engl. [doːtɚ] und dt. [tʰɔxtʰɐ]: Zunächst werden die beiden IPA-kodierten Sequenzen tokenisiert, d.h. in lautlich distinkte Segmente zerlegt, was wichtig ist, da die IPA-Kodierung bestimmte Laute mit Hilfe diakritischer Zeichen angibt. In einem weiteren Schritt werden die Segmente in Lautklassen umgewandelt und anschließend aliniert. Die alinierten Sequenzen werden schließlich wieder in ihr Ausgangsformat überführt.

4.3 Sequenz-, profil- und trigram-basierte multiple Alinierung

Die Sequenz- und profilbasierte Alinierung wurde ähnlich wie die paarweise Alinierung mit Rückgriff auf vorgefertigte biologische Python-Bibliotheken implementiert. Da hier jedoch keine für linguistische Zwecke geeigneten allgemeinen Module zur multiplen Alinierung bereitgestellt sind, wurde der Kode für die progressive, die profil- und die trigram-basierte Alinierung eigenhändig in Python geschrieben, wobei für die Schritte, die auf paarweiser Alinierung beruhen, die vorgefertigten Algorithmen der biologischen Bibliotheken verwendet wurden.

5 Ausblick

5.1 Erweiterung des lautklassenbasierten Ansatzes

Die derzeitige Fassung des lautklassenbasierten Ansatzes ist noch in ihrem Anfangsstadium begriffen. Es bedarf vor allem einer gezielteren empirischen Unterfütterung mit Rückgriff auf anerkannte und häufige Lautwandelprozesse, um die Lautklassen zu verfeinern und zu einem stochastisch fundierteren Ansatz zu gelangen.

5.2 Erweiterung der Scoring-Funktionen für Lautklassen, Trigramme und Profile

Die derzeitige Fassung der Scoring-Funktion für lautklassenbasierte Alinierungen ist weitestgehend provisorischer Natur und beruht vor allem auf meiner persönlichen Erfahrung und Intuition im Bereich der historischen Linguistik. Um eine tatsächliche stochastisch fundierte Heuristik zu erarbeiten, bedarf es einer gezielten Sichtung von Daten aus verschiedenen Sprachen. Gleichzeitig muss geklärt werden, inwieweit die für die Linguistik bedeutsame Einbeziehung kontextueller Informationen im Rahmen der gängigen Algorithmen realisiert werden kann. Die bisherigen Versuche, die auf einer einfachen Addition der Kosten für alle Segmente eines Trigramms mit geringerer Gewichtung der Links- und der Rechtsumgebung beruhen, haben bis her keine signifikanten Verbesserungen der Alinierungen gezeigt. Hier bedarf es neben einer theoretischen Fundierung insbesondere auch einer Ausweitung der Testbasis. In Bezug auf die Profil-Profil-Alinierung müssen erweiterte Scoring-Funktionen getestet werden, da die derzeit verwendete *sum of pairs* Bewertung gewisse Probleme aufweist, da hier durch die Berechnung des Durchschnitts gewisse Segmente überbewertet werden können, weshalb es ratsam scheint, die Scoring-Funktion um eine Komponente zu erweitern, bei der die Sequenzen anhand ihrer Position im Leitbaum unterschiedlich gewichtet werden (Möglichkeiten zur Implementierung dieses Verfahren werden u.a. in Thompson *et al.* 1994, 4676 beschrieben).

5.3 Erstellung umfangreicher Testsets

Bisher wurden die Algorithmen nur an kleinen Datensets getestet. Um eine realistische Aussage über deren Wirksamkeit machen zu können, müssen umfangreiche Testsets für verschiedene Sprachfamilien erstellt werden. Die Frage, inwiefern Sprachwandelprozesse überhaupt in einem universellen stochastischen Modell erfasst werden können, muss dabei zunächst offen bleiben. Für die Untersuchungen sollte jedoch die Annahme einer prinzipiellen Universalität zumindest bestimmter Lautwandelprozesse nicht von vornherein verworfen werden, da sich nur mit Rückgriff auf diese Annahme ein heuristisches, stochastisch basiertes Vorgehen gegenüber deterministischen Verfahren rechtfertigen lässt.

5.4 Erweiterung der Methoden zur multiplen Alinierung

Die multiple Alinierung, welche derzeit von dem Programm realisiert wird, liegt in ihren Möglichkeiten derzeit noch weit hinter den gängigen Algorithmen zurück, die in der Biologie entwickelt wurden. Daher sollten erweiterte Algorithmen der multiplen Alinierung für das Programm implementiert werden, um ihre Wirksamkeit in Bezug auf sprachliche Daten zu testen. Als Alternative zur progressiven multiplen Alinierung sind zwei Verfahren interessant, die in nächster Zeit implementiert werden sollen:

- **Iterative Verfahren:** Während progressive Verfahren auf einem Leitbaum aufbauen und einmal der Gesamtalinierung hinzugefügte Sequenzen nicht weiter verändern, erlauben iterative Verfahren die stete Neuberechnung der Alinierung, indem einmal alinierte Sequenzen aus der Gesamtalinierung herausgenommen und schrittweise wieder hinzugefügt werden (vgl. Durbin 2002, 148f). Ein relativ einfach zu realisierendes iteratives Verfahren stellt der Algorithmus von Barton & Sternberg (1987) dar, der als nächstes implementiert werden soll, um seine Leistungsfähigkeit gegenüber progressiven Alinierungsverfahren zu testen.

- **Hidden Markov Modelle:** Hidden Markov Modelle können verstanden werden als “abstract machine that has an ability to produce some output using coin tossing” (Jones & Pevzner 2004, 390).

> Zu Beginn befindet sich die Maschine in einer Reihe verborgener Zustände, basierend auf denen bestimmte Symbole zufällig ausgegeben werden, wobei die Ausgabe von Symbolen durch Wahrscheinlichkeiten beeinflusst werden kann. Da die verborgenen Zustände nicht bekannt sind, sondern nur die Ausgabe, ist es das Ziel des Beobachters, durch die Analyse der Ausgabesymbole die wahrscheinlichste Kette von Ausgangszuständen zu ermitteln (vgl. Jones & Pevzner 2004, 390). Wie genau mit Hilfe von Hidden Markov Modellen Alinierungen realisiert werden, kann an dieser Stelle noch nicht eindeutig dargestellt werden und bedarf der genaueren Untersuchung. Der Vorteil der Verwendung von Hidden Markov Modellen liegt jedoch darin, dass die für die Profil-Profil-Alinierung problematische Scoring-Funktion, die auf der *sum of pairs* beruht, durch Hidden Markov Modelle ersetzt werden kann. Ferner können Hidden Markov Modelle an alinierten und nicht-alinierten Testdaten trainiert werden (vgl. Durbin 2002, 149).

Abgesehen von den beiden erwähnten Verfahren gibt es in der Biologie eine Vielzahl weiterer Ansätze für multiple Alinierung (vgl. bspw. den Überblick in Gusfield 1997, 359-366). Ob und inwiefern diese für eine Anwendung in der historischen Linguistik sinnvoll sind, muss zunächst durch eine grobe Sichtung der Literatur untersucht und eventuell durch eine eigene Implementierung getestet werden.

5.5 Untersuchung von Ähnlichkeiten und Unterschieden in biologischen und linguistischen Verfahren

Die bisherigen Untersuchungen haben gezeigt, dass es eine Reihe von Parallelen zwischen biologischen und linguistischen Sequenzen gibt, die eine Übertragung der biologischen Verfahren auf die Linguistik sinnvoll erscheinen lassen. Diese Parallelen müssen jedoch klarer herausgearbeitet werden, wobei auch auf die auffälligen Unterschiede hingewiesen werden sollte. Nur durch eine Herausarbeitung von Gemeinsamkeiten und Unterschieden in den beiden Disziplinen kann eine erfolgreiche Einbeziehung biologischer Methoden in die historische Linguistik gewährleistet werden. Hierzu sind zunächst die unterschiedlichen Konzepte von Ähnlichkeit in der historischen Linguistik und der Biologie genauer zu untersuchen. Ferner müssen die unterschiedlichen Mechanismen von Wandelprozessen genauer beleuchtet werden.

References

Anttila, Raimo. 1972. *An introduction to historical and comparative linguistics*. New York: Macmillan. 2. Aufl. u.d.T.: Anttila, Raimo: Historical and comparative linguistics.

Barton, Geoffrey J., & Michael J. E. Sternberg. 1987. A strategy for the rapid multiple alignment of protein sequences : Confidence levels from tertiary structure comparisons. *Journal of Molecular Biology* 198.327 – 337.

Burlak, Svetlana Anatol'evna, & Sergej Anatol'evic Starostin. 2005. *Sravnitel'no-istoričeskoe jazykoznanie (Comparative-historical linguistics)*. Moskva: Akademia.

Cock, P J, T Antao, J T Chang, B A Chapman, C J Cox, A Dalke, I Friedberg, T Hamelryck, F Kauff, B Wilczynski, & M J de Hoon. 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25.1422–1423.

Damerau, Fred J. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM* 7.171–176.

Dolgopolsky, A. B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of northern eurasia. In *Typology Relationship and Time*, ed. by T. L. Shevoroshkin, Vitaly V.; Markey, Notes on Linguistics, 27–50. Karoma Publisher, Inc. Originally published in Russian as "Gipoteza drevnejščego rodstva jazykov Severnoj Evrazii (problemy fonetičeskich sootvetstvij)" in 1964.

Downey, Sean S., Brian Hallmark, Murray P. Cox, Peter Norquest, & Stephen Lansing. 2008. Computational feature-sensitive reconstruction of language relationships: Developing the aline distance for comparative historical linguistic reconstruction. *Journal of Quantitative Linguistics* 15.340–369.

Durbin, Richard. 2002. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press, 7th print edition.

Edgar, Robert C., & Kimmen Sjolander. 2004. A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics* 20.1301–1308.

Feng, D. F., & R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* 25.351–360.

Fox, Anthony. 1995. *Linguistic reconstruction: An introduction to theory and method*. Oxford University Press.

Gotoh, Osamu. 1982. An improved algorithm for matching biological sequences. *Journal of Molecular Biology* 162.705 – 708.

Gusfield, Dan. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge: Cambridge University Press.

Heeringa, Wilbert J., Peter Kleiweg, Charlotte Gooskens, & John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*, ed. by John Nerbonne & E. Hinrichs, 51–62, Sydney.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, & Dik Bakker. 2008a. Advances in automated language classification. In *Quantitative Investigations in Theoretical Linguistics*, ed. by Antti Arppe, Kaius Sinnemäki, & Urpu Nikann, 40–43. Helsinki: University of Helsinki.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, Andr´e Müller, & Dik Bakker. 2008b. Explorations in automated lexicostatistics. *Folia Linguistica* 20.116–121.

Jones, Neil C., & Pavel A. Pevzner. 2004. *An introduction to bioinformatics algorithms*. Cambridge, London: MIT Press.

Knight, Rob, Peter Maxwell, Amanda Birmingham, Jason Carnes, J Gregory Caporaso, Brett Easton, Michael Eaton, Micah Hamady, Helen Lindsay, Zongzhi Liu, Catherine Lozupone, Daniel McDonald, Michael Robeson, Raymond Sammut, Sandra Smit, Matthew Wakefield, Jeremy Widmann, Shandy Wikman, Stephanie Wilson, Hua Ying, & Gavin Huttley. 2007. Pycogent: a toolkit for making sense from sequence. *Genome Biology* 8.R171.

Kondrak, Grzegorz, 2002. *Algorithms for language reconstruction*. Toronto: University of Toronto dissertation.

Lass, Roger. 1997. *Historical linguistics and language change*. Cambridge: Cambridge University Press.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10.707–710.

Needleman, Saul B., & Christan D. Wunsch. 1970. A gene method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48.443–453.

Oommen, B. John. 1995. String alignment with substitution, insertion, deletion, squashing, and expansion operations. *Inf. Sci. Inf. Comput. Sci.* 83.89–107.

Rauhut, Reinhard. 2001. *Bioinformatik. Sequenz-Struktur-Funktion*. Weinheim, New York: Wiley-VCH.

Schwink, Frederick. 1994. *Linguistic typology, universality and the realism of reconstruction*. Washington: Institute for the Study of Man.

Serva, Maurizio, & Filippo Petroni. 2008. Indo-european languages tree by levenshtein distance. *EPL* 81.

Smith, T. F., & M. S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 1.195–197.

Thompson, J. D., D. G. Higgins, & T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22.4673–4680.

Wagner, Robert A., & Michael J. Fischer. 1974. The string-to-string correction problem. *J. ACM* 21.168–173.