Distanz- und Alignmentanalysen in der historischen Linguistik

Johann-Mattis List
Heinrich Heine Universität Düsseldorf

22. Januar 2010

Gliederung

Grundlegende Anmerkungen zur Einführung

Mengen- vs. Sequenzvergleiche

Mengenvergleiche Vergleich einer Anzahl ungeordneter, distinkter Elemente

Sequenzvergleiche Vergleich einer Anzahl geordneter Elemente, deren Distinktivität erst durch die Anordnung hergestellt wird

Sequenzvergleiche setzen eine *Alinierung* der Sequenzen voraus, da Sequenzdistanzen nur ermittelt werden können, wenn die korrespondierenden Segmente bestimmt wurden.

Unigrammatische und *n*-grammatische Segmentierung

Monogrammatisch	θ	i	ɣ	a	t	ɛ	r	a	
Bigrammatische	-θ	θi	iɣ	ɣa	at	tɛ	ɛr	ra	a-
Trigrammatische	--θ	-θi	θiɣ	iɣa	ɣat	atɛ	tɛr	ɛr-	r--

Tabelle: Mono- bi und trigram-basierte Segmentierung

Paarweise und multiple Alinierung

d	ɔː	-	-	t^h	ɚ	-	-
θ	i	ɣ	a	t	ɛ	r	a
t^h	ɔ	x	-	t^h	ɐ	-	-

Tabelle: Multiple Alinierung von Sequenzen

Grundidee des Algorithmus

- Erstellen einer Matrix mit allen möglichen Segmententsprechungen zweier Sequenzen
- Festsetzen von Kosten für die Gegenüberstellung der jeweiligen Segmente durch eine Vergleichsfunktion
  - Gegenüberstellung von Segmenten (Substitution & Match)
  - Einfügen von (Null)-Segmenten (Insertion & Deletion)
- Kumulative Aufrechnung von Kosten für alle möglichen Wege durch die Matrix

Die Vergleichsfunktion der Levenshtein-Distanz

Entscheidung	**Bedingung**	**Kosten**
Gegenüberstellung	Identität von Segmenten	0
	Verschiedenheit von Segmenten	1
Einfügen und Ersetzen		1

Tabelle: Die Vergleichsfunktion der Levenshtein-Distanz

Erstellen der Matrix

		h	e	r	z
	-	-	-	-	-
h	-	h	h	h	h
e	-	h	e	e	e
a	-	h	e	r	a
r	-	h	e	r	r
t	-	h	e	r	z

Tabelle: Vergleichsmatrix für den Wagner-Fischer-Algorithmus

Berechnen der Kosten für jeden Pfad

0 -/-	1 -/h	2 -/e	3 -/r	4 -/z
1 h/-	0 h/h	1 -/e	2 -/r	3 -/z
2 e/-	1 e/-	0 e/e	1 -/r	2 -/z
3 a/-	2 a/-	1 a/-	1 a/r	2 -/z
4 r/-	3 r/-	2 a/-	1 r/r	2 -/z
5 t/-	4 t/-	3 a/-	2 t/r	2 t/z

Tabelle: Vergleichsmatrix nach der Auswertung (mit Kosten)

Transpositionen

Levenshtein	f	r	o	m	a	g	e	-	-
	f	o	r	m	a	g	g	i	o
Damerau-Levenshtein	f	ro		m	a	g	e	-	-
	f	or		m	a	g	g	i	o

Tabelle: Levenshtein und Damerau-Levenshtein

Konsekutive Indels

Traditionell	d	i	d	oː	m	i
	d	a	-	-	m	-
Gotoh-Erweiterung	d	i	d	oː	m	i
	-	-	d	a	m	-

Tabelle: Konsekutive *indels* in der Sequenzalinierung

Kompressionen und Expansionen

Levenshtein	d	ɔ	-	t^h	ɐ
	t^h	ɔ	x	t^h	ɐ
Oommen-Erweiterung	d	ɔ	t^h		ɐ
	t^h	ɔ	xt^h		ɐ

Tabelle: Kompression und Expansion in der Sequenzalinierung

Lokale Alinierung

Levenshtein	-	ā	p	a	k	o	s	ī	s	-
	w	ā	p	i	k	o	n	ō	h	a
Lokale Alinierung		ā	p	a	k	o	sīs			
	w	ā	p	i	k	o	nōha			

Probleme beim Alinieren phonetischer Daten

θ	i	ɣ	a	t	ɛ	r	a
d	ɔː	t^h	ɔ̯	-	-	-	-

Tabelle: Problem der Alinierung phonetischer Sequenzen

Wie gut muss die Vergleichsfunktion sein?

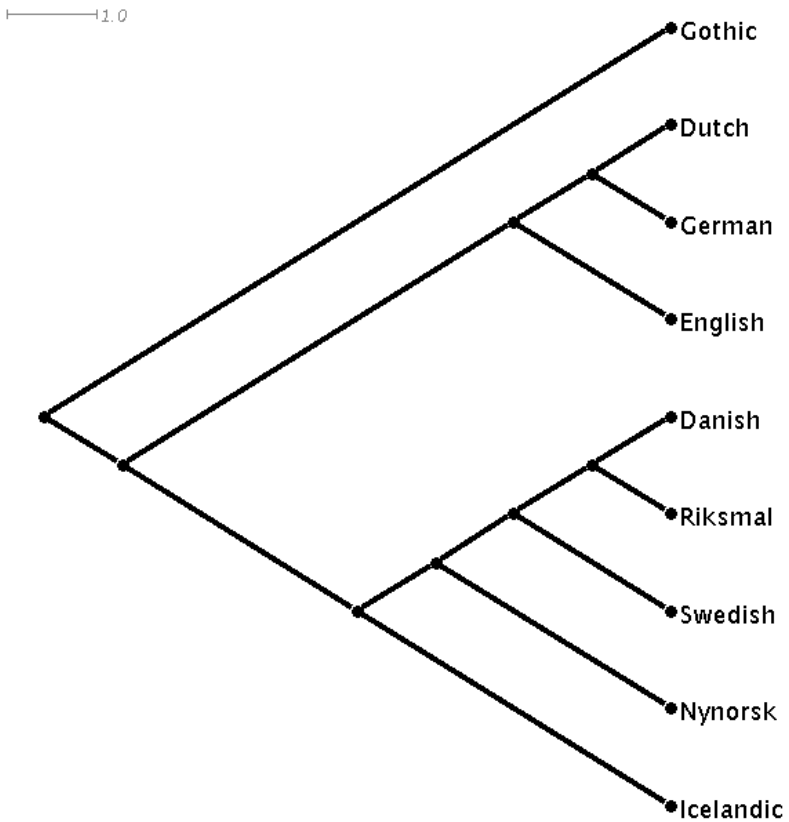

Abbildung: Neighbor-Analyse von Levenshtein-Distanzen

Lautklassen in der Alinierung

- Zuweisung von Segmenten zu einheitlichen Klassen, innerhalb derer Lautwandel als wahrscheinlicher angenommen wird, als außerhalb der Klassen
- Umwandlung der Segmente in ihre jeweiligen Klassen
- Durchführung “traditioneller” Alignmentanalysen

Der Vorteil von lautklassenbasierten Ansätzen liegt in ihrer leicht zu realisierenden Implementierung. Ferner können phonetisch unzureichende Sprachdaten meist relativ einfach in ein Klassenformat überführt werden.

Dolgopolskys Lautklassen

No.	Typ	Beschreibung	Bsp.
1	P	labiale Obstruenten	p,b,f
2	T	dentale Obstruenten	d,t,θ,ð
3	S	alveolare, postalveolare und retroflexe Frikative	s,z,ʃ,ʒ
4	K	velare und postvelare Obstruenten und Affrikaten	k,g,ts,tʃ
5	M	labialer Nasal	m
6	N	übrige Nasale	n,ɲ,ŋ
7	R	Trills, Taps, Flaps und laterale Approximanten	r,l
8	W	stimmhafter labialer Frikativ und initiale gerundete Vokale	v,u
9	J	palataler Approximant	j
10	ø	Laryngale und initialer velarer Nasal	h,ɦ,ŋ

Tabelle: Dolgopolskys Klassifizierung von Lautwandeltypen

Alinierung mit Hilfe von Dolgopolskys Lautklassen

Interne Alinierung	T	V	K	V	T	V	R	V
	T	V	-	-	T	V	-	-
Ausgabe	θ	i	ɣ	a	t	ɛ	r	a
	d	ɔː	-	-	t^h	ɚ	-	-

Tabelle: Interne und externe Darstellung der Dolgopolsky-Alinierung

The Automated Similarity Judgment Program

Sprache	Bedeutung	IPA	ASJP-Code
engl.	'das'	ðis	8is
engl.	'Mund'	mauθ	mau8
engl.	'Zunge'	t^həŋ	th~3N
dt.	'Fisch'	fiʃ	fiS
engl.	'Zahn'	tuːθ	tu8
dt.	'Schwester'	ʃwɛst^{h}ɐ	Swasth~a

Tabelle: Das universale Alphabet des ASJP-Projektes

Covington

Penalty	Conditions
0	Exact match of consonants or glides (w, y)
5	Exact match of vowels
10	Match of two vowels that differ only in length, or i and y, or u and w
30	Match of two dissimilar vowels
60	Match of two dissimilar consonants
100	Match of two segments with no discernible similarity
40	Skip preceded by another skip in the same word
50	Skip not preceded by another skip in the same word

Tabelle: Covingtons 'Evaluationsmetrik' für die Alinierung

Kondraks ALINE

Syllabic	5	Place	40
Voice	10	Nasel	10
Lateral	10	Aspirated	5
High	5	Back	5
Manner	50	Retroflex	10
Long	1	Round	5

Tabelle: Merkmalssalienzen in ALINE

Arbeitsstand und Ausblick

Python-Programm zur umfangreichen Sequenzanalyse

- Modularer Aufbau, Module für Sequenzanalysen (Alignments und Distanzberechnungen), phonetische Analysen, automatische Kognatenerkennung
- Eingabe in Form Unicode-kodierter csv-Dateien (comma separated value), deren Struktur sich am Format etymologischer Wörterbucher in der STARLING-Software orientiert
- Ausgabe von Analysen und Berechnungen in verschiedenen Formaten, die wahlweise durch phylogenetische Softwarepakete weiterverarbeitet werden können (Phylip, Nexus)
- Skripte für Daten-Ein- und -Ausgabe sind bereits realisiert, ferner sind eine Reihe von Algorithmen/Verfahren zur Sequenzanalyse (erweiterter DPA, Dolgopolsky, ASJP, Covington) bereits verfügbar

Das war's!