Phonetic Alignment Based on Sound-Classes

A New Method for Sequence Comparison in Historical Linguistics

Johann-Mattis List*

*Institute for Romance Languages and Literature
Heinrich Heine University Düsseldorf

ESSLLI 2010 Students' Session

Structure of the Talk

Introduction
- Sequence Comparison in Historical Linguistics
- Alignment Analyses in Historical Linguistics

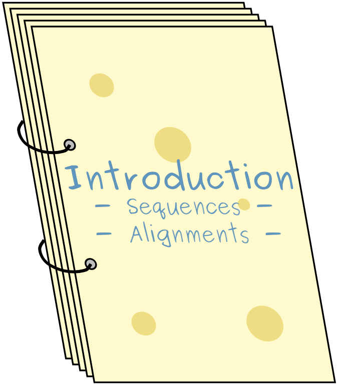
Basic Procedures for Automatic Alignment Analyses
- The Dynamic Programming Algorithm
- Multiple Sequence Alignment

Sound Classes in Historical Linguistics
- Two Perspectives on Similarity in Linguistics
- The Conception of Sound Classes

The Python Library for Sound-Class Based Alignment
- General Working Principle
- Pairwise and Multiple Alignments

Performance of the Method
- Pairwise Alignments
- Multiple Alignments

Introduction



HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

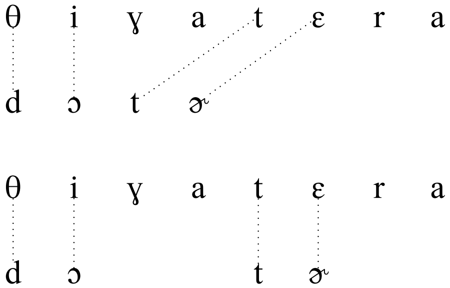
Sequence Comparison in Historical Linguistics

- Basic of the comparative method
- Basic of the detection of regular sound correspondences
- Basic of the proof of genetic relationship
- Basic of genetic language classification

Alignment Analyses in Historical Linguistics

- Sequences – in contrast to sets – consist of non-unique elements which retrieve distinctive function only because of their order.
- In alignment analyses, the corresponding elements of two or more sequences are ordered in such a way that they are set against each other.
- Sequence comparison in historical linguistics is always based on phonetic alignment.

Alignment Analyses in Historical Linguistics



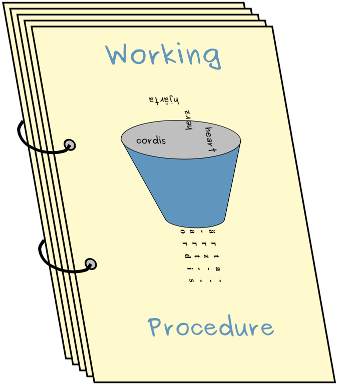
θ i ɣ a t ɛ r a

d ɔ t ɚ

θ i ɣ a t ɛ r a

d ɔ t ɚ

Basic Procedures for Automatic Alignment Analyses



The Dynamic Programming Algorithm

- Create a matrix which confronts all segments of the sequences under comparison, either with each other, or with alternative null-sequences (fills).
- Seek the path through the matrix which is of the lowest general costs.
- Calculate the costs cumulatively by means of a specific scoring function that penalizes the matching of segments with each other and likewise the insertion and deletion of segments in any of the sequences.

The Dynamic Programming Algorithm

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

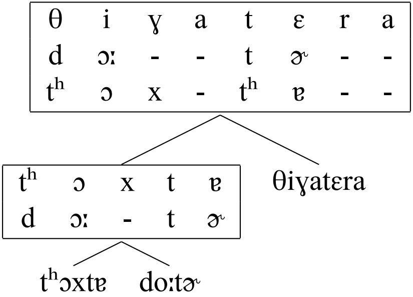
The Dynamic Programming Algorithm

0	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	1	1	2
4	3	2	2	2	2

Multiple Sequence Alignment: Guide-Tree Heuristics

- Due to computational restrictions, multiple sequence alignment (MSA) is based on heuristics.
- Heuristics based on guide-trees are the most common ones used in computational biology.
- Based on pairwise alignment scores, a guide-tree is reconstructed, and the sequences are stepwise added to the MSA along it (Feng & Dolittle 1987).

Multiple Sequence Alignment: Guide-Tree Heuristics



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

tʰ	ɔ	x	t	ɐ
d	ɔː	-	t	ɚ

θiɣatɛra

tʰɔxtɐ

doːtɚ

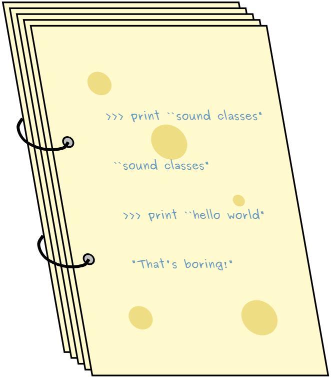
Multiple Sequence Alignment: Profiles

- The guide-tree heuristic can be enhanced by the application of profiles.
- A profile consists of the relative frequency of all segments of a MSA in all its positions, thus, a profile represents a MSA as a sequence of vectors.
- Aligning profiles to profiles instead of aligning two representative sequences of two given MSA yields better results, since more information can be taken into account.

Multiple Sequence Alignment: Profiles

	Multiple Alignment: Traditional Format						
	ʧ	-	l	o	vʲ	ɛ	k
	ʧ	-	-	o	v	ɛ	k
	ʧʲ	ɪ	l	ɐ	vʲ	ɛ	k
	ʧ	-	w	ɔ	vʲ	ɛ	k
	Multiple Alignment: Profile Representation						
ʧ	.75						
ʧʲ	.25						
l			.50				
o				.50			
v					.25		
vʲ					.75		
ɐ				.25			
ɛ						1.0	
ɪ		.25					
k							1.0
w			.25				
ɔ				.25			
-		.75	.25				

Sound Classes in Historical Linguistics



```
>>> print ``sound classes"

``sound classes"

>>> print ``hello world"

"That's boring!"
```

Two Perspectives on Similarity in Linguistics

Synchronic Similarity Sounds in different languages are judged to be similar, if they show resemblences regarding the way they are produced or perceived.

Diachronic Similarity Sounds in different languages are judged to be similar, if they go back to a common ancestor.

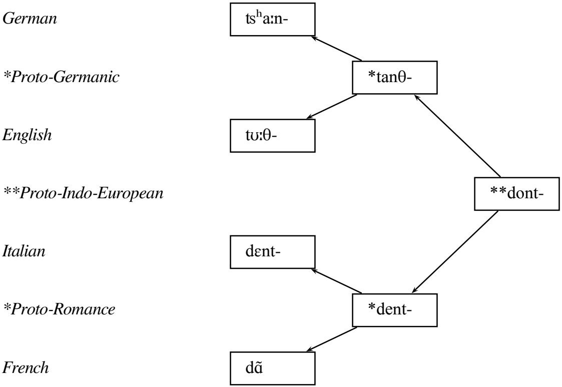
Two Perspectives on Similarity in Linguistics

Language	Word	Meaning
Mandarin	$\mathrm{ma}_{55}\mathrm{ma}_{3}$	“mother”
German	mama	“mother”
Russian	tak	“in this way”
German	t^{h}aːk	“day”

Two Perspectives on Similarity in Linguistics

Language	Word	Meaning
German	ʦʰaːn	“tooth”
English	tuːθ	“tooth”
Italian	dɛntɛ	“tooth”
French	dɑ̃	“tooth”

Two Perspectives on Similarity in Linguistics



German ts^{h}aːn-

**Proto-Germanic* *tanθ-

English tʊːθ-

***Proto-Indo-European* **dont-

Italian dɛnt-

**Proto-Romance* *dent-

French dɑ̃

The Conception of Sound Classes

Key Assumption of the Sound Class Approach

It is possible “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).

A Diachronic Definition of Similarity

Similarity is not based on synchronic resemblances of sounds but on on class-membership: two sounds, how dissimilar they may be from a synchronic perspective, may still belong to the same class.

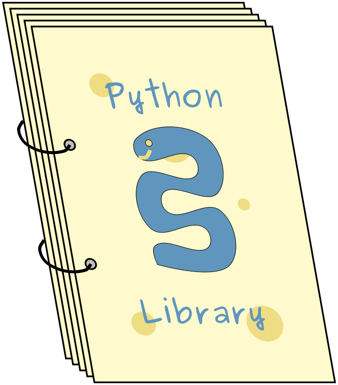
The Conception of Sound Classes

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

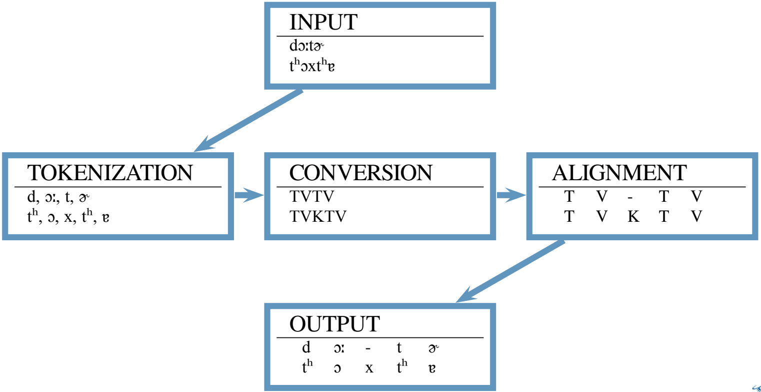
Table: Dolgopolsky’s (1986) Sound Classes

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

The Python Library for Sound-Class-Based Alignment



General Working Principle



INPUT

dɔːtɚ
tʰɔxtʰɐ

TOKENIZATION

d, ɔː, t, ɚ
tʰ, ɔ, x, tʰ, ɐ

CONVERSION

TVTV
TVKTV

ALIGNMENT

T V - T V
T V K T V

OUTPUT

d ɔː - t ɚ
tʰ ɔ x tʰ ɐ

Pairwise and Multiple Alignments

Pairwise Alignments

- Based on pairwise2 of BioPython (Cock et al. 2009)
- Scoring functions adapted for Dolgopolsky sound classes
- Global and local alignment analyses

Multiple Alignments

- MSA based on guide-trees (Feng & Doolittle 1987)
- MSA based on profiles (Thompson et al. 1994)
- Guide-trees calculated with PyCogent (Knight et al. 2007)
- Scoring function based on sum of pairs (Durbin 2002: 139f)

Performance of the Method

w	o	l	-	d	e	m	o	r	t
w	-	l	a	d	i	m	i	r	-
w	a	l	-	d	e	m	a	r	-

Pairwise Alignments: Covington's (1996) Testset

Sound Classes vs. ALINE (Kondrak 2002)

Identical results:	71 / 82 cases
Double outputs where ALINE has one output:	6 cases
Double outputs matching ALINE's single output:	4 cases
Double outputs superior to ALINE:	1 case
Double outputs both fail:	1 case
ALINE superior to Sound Classes:	3 cases
Sound Classes superior to ALINE:	2 cases

Pairwise Alignments: Examples

	Sound-Class-Approach			ALINE	
1	Engl. daughter / Old Grk. θυγατήρ “daughter”				
	d o - - t ə r tʰ u g a t eː r		 tʰu	d o t ə r g a t eː r	
2	Engl. this / Grm. dieses “this”				
	ð i s d iː z	 əs	 diː	ð i z z ə s	
3	Engl. tooth / Lat. dentis “tooth”				
	t u - θ d e n t	 is	 den	t u θ t i s	

Multiple Alignments: First Tests on Small Samples

Simple Guide-Tree-Based MSA

t^h	u	g	a	t	eː	r
t	o	x	-	t	ə	r
d	o	-	-	t	ə	r
d	u	-	ʃ	t	i	-
d	u	h	i	t	aː	r

Profile-based MSA

t^h	u	g	a	t	eː	r
t	o	x	-	t	ə	r
d	o	-	-	t	ə	r
d	u	ʃ	-	t	i	-
d	u	h	i	t	aː	r

Old Grk. θυγατήρ / Grm. Tochter / Engl. daughter / OCS дъщи / Skr. duhitār “daughter”

Multiple Alignments: First Tests on Small Samples

Simple Guide-Tree-Based MSA

ʧ	-	l	o	vʲ	ɛ	k
ʧ	-	-	o	v	ɛ	k
ʧʲ	ɪ	l	ɐ	vʲ	ɛ	k
ʧ	w	-	ɔ	vʲ	ɛ	k

Profile-based MSA

ʧ	-	l	o	vʲ	ɛ	k
ʧ	-	-	o	v	ɛ	k
ʧʲ	ɪ	l	ɐ	vʲ	ɛ	k
ʧ	-	w	ɔ	vʲ	ɛ	k

Czech člověk / Bulgarian човек / Russian человек / Polish człowiek “human”

Thank You
for
listening!

Special thanks to:
Shiju Lal NS
Ovidiu Popa
Tal Dagan
Hans Geisler