

# Improving Phonetic Alignment by Handling Secondary Sequence Structures

Johann-Mattis List\*

\* Institute for Romance Languages and Literature  
Heinrich Heine University Düsseldorf

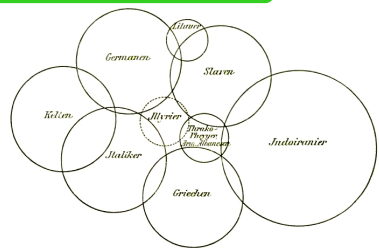
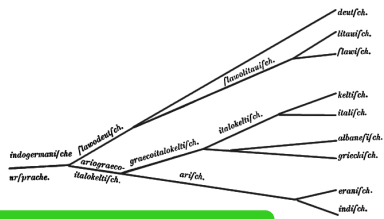
2012/08/10

# Structure of the Talk

- 1 Historical Linguistics
  - Keys to the Past
  - Comparative Method
  - Sound Correspondences
- 2 Sequence Comparison
  - Sequences
  - Alignment Analyses
  - Alignment Modes
- 3 Secondary Alignment
  - Secondary Sequence Structures
  - Secondary Alignment Problem
  - Secondary Alignment Algorithm
- 4 Phonetic Alignment
  - SCA
  - Paradigmatic Aspects
  - Syntagmatic Aspects
- 5 Evaluation
  - Evaluation Measures
  - Gold Standard
  - Results

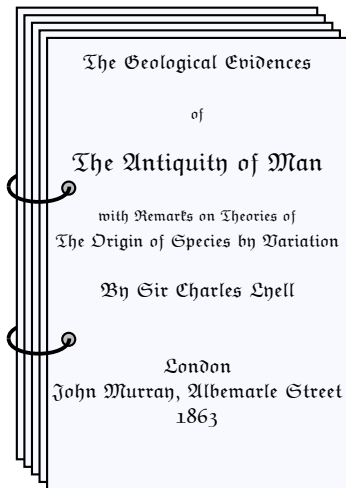


# Historical Linguistics

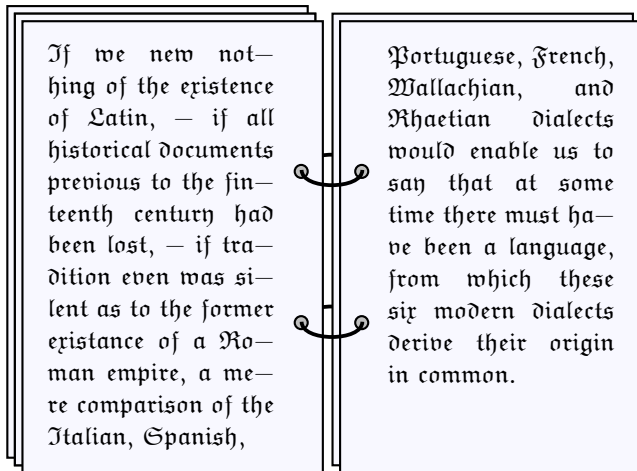


# Charles Lyell on Languages

# Charles Lyell on Languages



## Charles Lyell on Languages



# Historical Scenarios

*German*

ts    a:    n

*English*

t    u:    θ

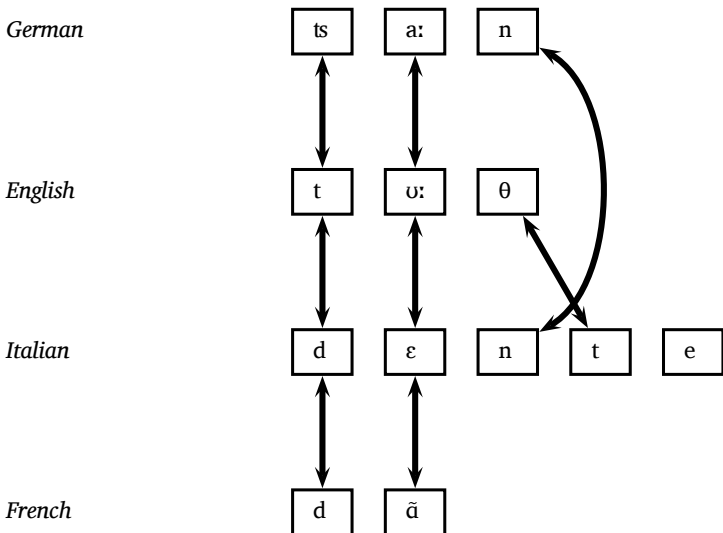
*Italian*

d    ε    n    t    e

*French*

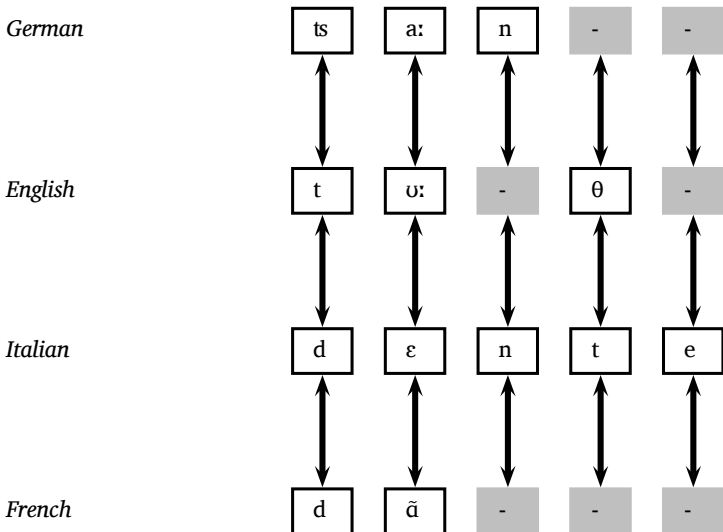
d    ã

# Historical Scenarios

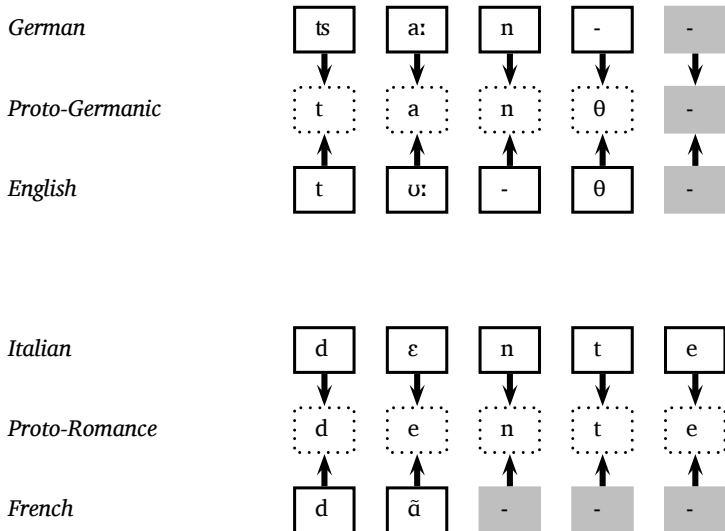




# Historical Scenarios



# Historical Scenarios



# Historical Scenarios

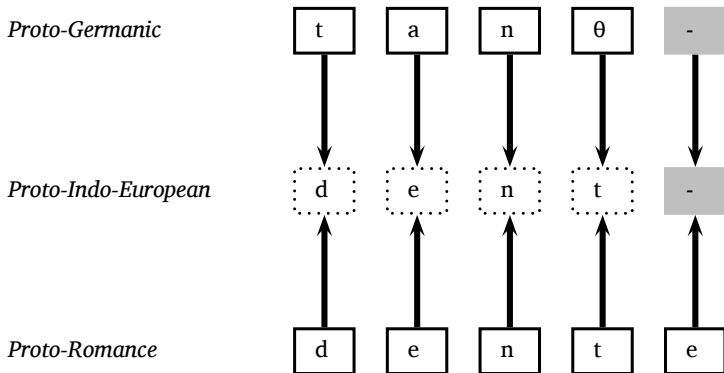
*Proto-Germanic*

t a n θ -

*Proto-Romance*

d e n t e

# Historical Scenarios



# Historical Scenarios

*Proto-Indo-European*

d

e

n

t



# The Comparative Method

- Compile an initial list of putative cognate sets.
- Extract an initial list of putative sets of sound correspondences from the initial cognate list.
- Refine the cognate list and the correspondence list by
  - adding and deleting cognate sets from the cognate list, depending on whether they are consistent with the correspondence list or not, and
  - adding and deleting correspondence sets from the correspondence list, depending on whether they are consistent with the cognate list or not.
- Finish when the results are satisfying enough.

# Sound Correspondences

- Sequence similarity is determined on the basis of systematic sound correspondences as opposed to similarity based on surface resemblances of phonetic segments.
- Lass (1997) calls this notion of similarity **phenotypic** as opposed to a **genotypic** notion of similarity.
- The most crucial aspect of correspondence-based similarity is that it is **language-specific**: Genotypic similarity is never defined in general terms but always with respect to the language systems which are being compared.



# Sound Correspondences

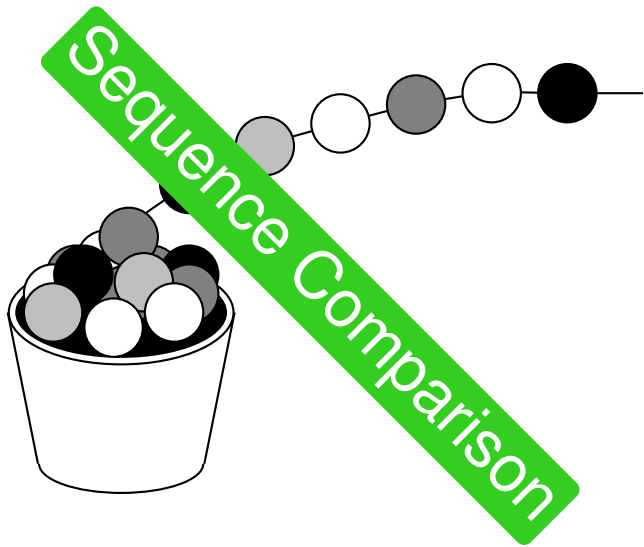
- Sequence similarity is determined on the basis of systematic sound correspondences as opposed to similarity based on surface resemblances of phonetic segments.
- Lass (1997) calls this notion of similarity **phenotypic** as opposed to a **genotypic** notion of similarity.
- The most crucial aspect of correspondence-based similarity is that it is **language-specific**: Genotypic similarity is never defined in general terms but always with respect to the language systems which are being compared.

Meaning	German	Dutch	English
“tooth”	<i>Zahn</i> [ts a:n]	<i>tand</i> [t ant]	<i>tooth</i> [t u:θ]
“ten”	<i>zehn</i> [ts e:n]	<i>tien</i> [t i:n]	<i>ten</i> [t ɛn]
“tongue”	<i>Zunge</i> [ts ʊŋə]	<i>tong</i> [t ɔŋ]	<i>tongue</i> [t ʌŋ]

# Sound Correspondences

- Sequence similarity is determined on the basis of systematic sound correspondences as opposed to similarity based on surface resemblances of phonetic segments.
- Lass (1997) calls this notion of similarity **phenotypic** as opposed to a **genotypic** notion of similarity.
- The most crucial aspect of correspondence-based similarity is that it is **language-specific**: Genotypic similarity is never defined in general terms but always with respect to the language systems which are being compared.

Meaning	Shanghai	Beijing	Guangzhou
“nine”	[tɕiŋ <sup>35</sup> ]	Beijing [tɕiou <sup>214</sup> ]	[kɛu <sup>35</sup> ]
“today”	[tɕiŋ <sup>55</sup> tsɔ <sup>21</sup> ]	Beijing [tɕiə <sup>55</sup> ]	[kɛm <sup>53</sup> jet <sup>2</sup> ]
“rooster”	[koŋ <sup>55</sup> tɕi <sup>21</sup> ]	Beijing [kuŋ <sup>55</sup> tɕi <sup>55</sup> ]	[kɛi <sup>55</sup> koŋ <sup>55</sup> ]

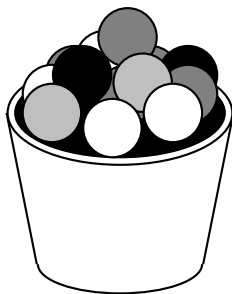


# Sequences

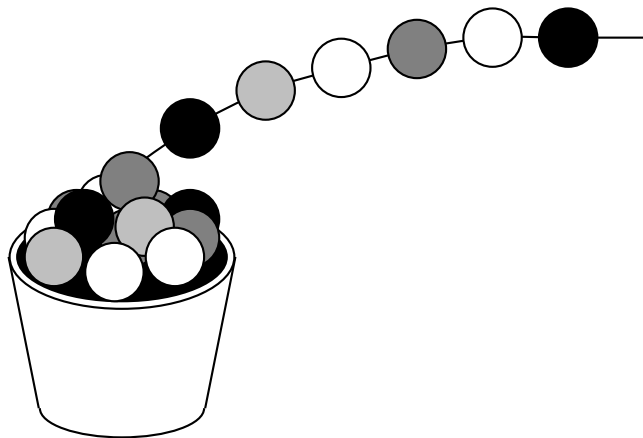
## Definition 1

Given an *alphabet* (a non-empty finite set, whose elements are called *characters*), a *sequence* is an ordered list of characters drawn from the alphabet. The elements of sequences are called *segments*. (cf. Böckenbauer & Bongartz 2003: 30f)

# Sequences



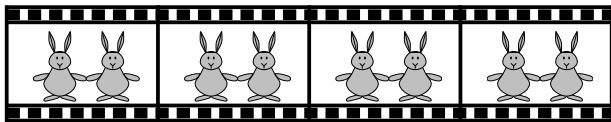
# Sequences



# Sequences



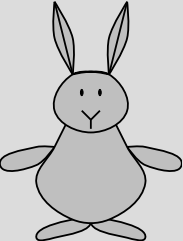
# Sequences





# Sequences

**Baked Rabbit**



1 rabbit  
1 1/2 tsp. salt  
1 1/8 1/8 tsp. pepper  
1 1/2 c. onion slices

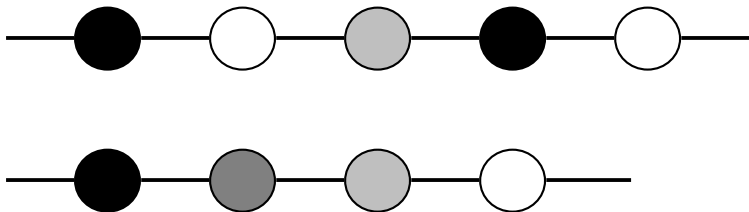
- Rub salt and pepper on rabbit pieces.
- Place on large sheet of aluminium foil.
- Place onion slices on rabbit.
- Bake at 350 degrees.
- Eat when done and tender.

# Alignment Analyses

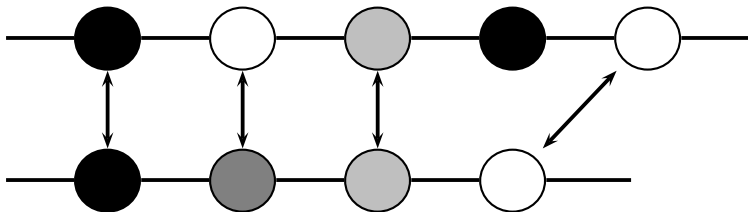
## Definition 2

An *alignment* of two sequences  $s$  and  $t$  is a two-row matrix in which both sequences are arranged in such a way that all matching and mismatching segments occur in the same column, while empty cells, resulting from empty matches, are filled with gap symbols. (cf. Kruskal 1983)

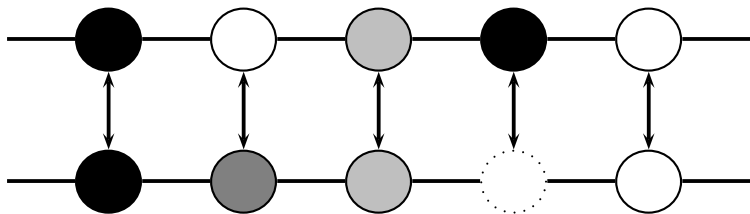
# Alignment Analyses



# Alignment Analyses



# Alignment Analyses



# Global Alignment

Global alignment analyses are the most basic way to compare sequences. The traditional Needleman-Wunsch algorithm (Needleman and Wunsch 1971) conducts global alignment analyses, and the Levenshtein distance (edit distance, Levenshtein 1965) is defined for global alignments.

# Global Alignment

Global alignment analyses are the most basic way to compare sequences. The traditional Needleman-Wunsch algorithm (Needleman and Wunsch 1971) conducts global alignment analyses, and the Levenshtein distance (edit distance, Levenshtein 1965) is defined for global alignments.

Mode	Alignment
global	G R E E N C A T F I S H H U N T E R A F A T C A T - - - H U N T E R

## Semi-Global Alignment

Semi-global alignment analyses do not necessarily compare two sequences as a whole but allow prefixes and suffixes to be ignored in an alignment analysis, if these would otherwise increase the cost of the optimal alignment. Computationally, this is done by setting the costs for gaps inserted in the begin and at the end of an alignment to zero.



# Semi-Global Alignment

Semi-global alignment analyses do not necessarily compare two sequences as a whole but allow prefixes and suffixes to be ignored in an alignment analysis, if these would otherwise increase the cost of the optimal alignment. Computationally, this is done by setting the costs for gaps inserted in the begin and at the end of an alignment to zero.

Mode	Alignment
<b>global</b>	G R E E N C A T F I S H H U N T E R A F A T C A T - - - H U N T E R
<b>semi-global</b>	G R E E N - C A T F I S H H U N T E R - - - - A F A T C A T H U N T E R

## Local Alignment

While semi-global alignment analyses allow prefixes and suffixes to be ignored only if one sequence contains a prefix or suffix while the other does **not**, local alignment analyses (Smith-Waterman algorithm, Smith and Waterman 1981) only align the best scoring **subsequences** of two sequences, while leaving the rest of the sequences completely unaligned. Computationally, this is done by prohibiting that the cost of an alignment analysis goes beyond zero.

# Local Alignment

While semi-global alignment analyses allow prefixes and suffixes to be ignored only if one sequence contains a prefix or suffix while the other does **not**, local alignment analyses (Smith-Waterman algorithm, Smith and Waterman 1981) only align the best scoring **subsequences** of two sequences, while leaving the rest of the sequences completely unaligned. Computationally, this is done by prohibiting that the cost of an alignment analysis goes beyond zero.

Mode	Alignment
<b>global</b>	G R E E N C A T F I S H H U N T E R A F A T C A T - - - H U N T E R
<b>semi-global</b>	G R E E N - C A T F I S H H U N T E R - - - - A F A T C A T H U N T E R
<b>local</b>	GREEN CATFISH   H U N T E R A FAT CAT   H U N T E R

## Diagonal Alignment

While local alignment analyses leave unalignable parts of sequences unaligned, diagonal alignment analyses (DIALIGN algorithm, Morgenstern 1996) align sequences globally, but search for local similarities at the same time. Local similarities are defined as “diagonals”, i.e. ungapped alignments. Diagonal alignment analyses maximize the score of diagonals in an alignment.

# Diagonal Alignment

While local alignment analyses leave unalignable parts of sequences unaligned, diagonal alignment analyses (DIALIGN algorithm, Morgenstern 1996) align sequences globally, but search for local similarities at the same time. Local similarities are defined as “diagonals”, i.e. ungapped alignments. Diagonal alignment analyses maximize the score of diagonals in an alignment.

Mode	Alignment
<b>global</b>	G R E E N C A T F I S H H U N T E R A F A T C A T - - - H U N T E R
<b>semi-global</b>	G R E E N - C A T F I S H H U N T E R - - - - A F A T C A T H U N T E R
<b>local</b>	GREEN CATFISH   H U N T E R A FAT CAT   H U N T E R
<b>diagonal</b>	- - - - G R E E N C A T F I S H H U N T E R A F A T - - - - C A T - - - H U N T E R

# Secondary Alignment

secondary sequence structures

secondary sequence structures

secondary sequence structures

secondary sequence structures

secondary sequence structures

SECONDARY SEQUENCE STRUCTURE

secondary sequence structures

secondary sequence structures

# Secondary Sequence Structures

Apart from a **primary** structure, sequences can also have a **secondary structure**. Primary structure refers to the order of segments. Secondary structure refers to the order of **secondary segments**, i.e. segments that result from the grouping of primary segments into higher units.

# Secondary Sequence Structures

Apart from a **primary** structure, sequences can also have a **secondary structure**. Primary structure refers to the order of segments. Secondary structure refers to the order of **secondary segments**, i.e. segments that result from the grouping of primary segments into higher units.

"ABCEFGIJK"

→ "ABC.EFG.IJK"



## Secondary Sequence Structures

Apart from a **primary** structure, sequences can also have a **secondary structure**. Primary structure refers to the order of segments. Secondary structure refers to the order of **secondary segments**, i.e. segments that result from the grouping of primary segments into higher units.

"ABCEFGIJK"

→ "ABC.EFG.IJK"

"THECATFISHHUNTER"

→ "THE.CATFISH.HUNTER"

# Secondary Sequence Structures

Apart from a **primary** structure, sequences can also have a **secondary structure**. Primary structure refers to the order of segments. Secondary structure refers to the order of **secondary segments**, i.e. segments that result from the grouping of primary segments into higher units.

"ABCEFGIJK"

→ "ABC.EFG.IJK"

"THECATFISHHUNTER"

→ "THE.CATFISH.HUNTER"

"KARAOKE"

→ "KA.RA.O.KE"

# The Secondary Alignment Problem

## Secondary Alignment Problem

Given two sequences  $s$  and  $t$  of length  $m$  and  $n$  which have the primary structures  $s_1, \dots, s_m$  and  $t_1, \dots, t_n$ , and the secondary structures  $s_{0 \rightarrow i}, \dots, s_{j \rightarrow m}$  and  $t_{0 \rightarrow k}, \dots, t_{l \rightarrow n}$ , find an alignment of maximal score in which segments belonging to the same secondary segment in  $s$  only correspond to segments belonging to the same secondary segment in  $t$ , and vice versa.

# The Secondary Alignment Problem

Mode	Alignment																			
global	T	H	E		C	A	T	-	F	I	S	H		H	U	N	T	S		
	T	H	E		C	A	T		F	I	S	H	-	E	-	-	-	S		
semiglobal	T	H	E		C	A	T	-	F	I	S	H	-	-	-	H	U	N	T	S
	T	H	E		C	A	T		F	I	S	H	E	S		-	-	-	-	-
local	T	H	E		C	A	T	-	F	I	S	H		HUNTS						
	T	H	E		C	A	T		F	I	S	H		ES						
diagonal	T	H	E		C	A	T	-	F	I	S	H	-		H	U	N	T	S	
	T	H	E		C	A	T		F	I	S	H	E	-	-	-	-	-	S	
secondary	T	H	E		C	A	T	F	I	S	H		H	U	N	T	-	S		
	T	H	E		C	A	T	-	-	-	-		F	I	S	H	E	S		

# A Secondary Alignment Algorithm

**Algorithm 1:** Secondary( $x, y, g, r, \text{score}$ )

**comment:** matrix construction and initialization

...

**comment:** main loop

**for**  $i \leftarrow 1$  **to** length( $x$ )

{	do	{	<b>for</b> $j \leftarrow 1$ <b>to</b> length( $y$ )	{	$M[i-1][j-1] + \text{score}(x_{i-1}, y_{j-1})$
			<b>comment:</b> check for restriction 2		<b>if</b> $x_{i-1} = r$ <b>and</b> $y_{j-1} \neq r$ <b>and</b> $j \neq \text{length}(y)$ <b>then</b> $-\infty$ <b>else</b> $M[i-1][j] + g$
{	do	{	$M[i][j] \leftarrow \max$	{	<b>if</b> $y_{j-1} = r$ <b>and</b> $x_{i-1} \neq r$ <b>and</b> $i \neq \text{length}(x)$ <b>then</b> $-\infty$ <b>else</b> $M[i][j-1] + g$

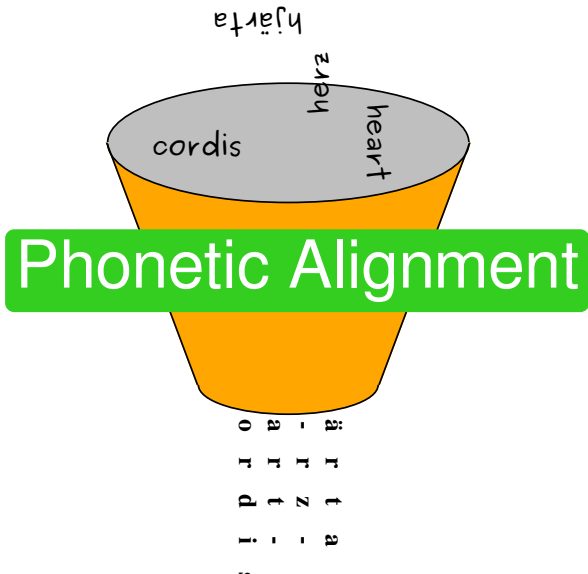
# A Secondary Alignment Algorithm

①		A	.	B	C	.	D	E
	0	-A	-2	-3	-4	-5	-6	-7
A	-1	A	0	-1	-2	-3	-4	-5
A	-2	A	0	-1	-2	-3	-4	-5
B	-3	B	-1	B	0	-1	-2	-3
C	-4	C	-2	C	0	2	-1	-1
D	-5	D	-3	D	-1	D	2	1
E	-6	E	-4	E	0	E	1	E
.	-7	.	-5	.	-3	.	-1	.
E	-8	E	-6	E	-4	E	0	E

②		A	.	B	C	.	D	E
	0	-A	-2	-3	-4	-5	-6	-7
A	-1	A	A	A	B	-C	A	A
A	-2	A	0	A	A	A	A	A
B	-3	B	-1	B	B	-C	B	B
C	-4	C	-2	C	C	C	C	C
D	-5	D	-3	D	D	D	D	D
E	-6	E	-4	E	E	E	E	E
.	-7	.	-5	.	-3	.	-1	.
E	-8	E	-6	E	-4	E	-4	E

# A Secondary Alignment Algorithm

- The extension for secondary alignment is independent of the underlying alignment mode.
- Global, semi-global, local, and diagonal alignment analyses that are sensitive for secondary sequence structures can be carried out.
- The only requirement of the algorithm in contrast to the traditional alignment algorithms is the boundary character which has to be specified by the user.





## Sound-Class-Based Phonetic Alignment (SCA)

- SCA (List 2012) is a new method for pairwise and multiple phonetic alignment, implemented as part of LingPy (<http://linguist.de/lingpy>), a Python library for quantitative tasks in historical linguistics.
- SCA is based on a novel framework for phonetic alignment that combines both the most recent developments in computational biology with new approaches to sequence modelling in historical linguistics and dialectology.
- According to the new framework for sequence modelling, sound sequences are internally represented in different layers which relate to both important paradigmatic and syntagmatic aspects of linguistic sequences.

# Sound Classes

## Sound Classes

Sounds which often occur in correspondence relations in genetically related languages can be clustered into classes (types). It is assumed “that phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’ ” (Dolgopolsky 1986: 35).

# Sound Classes

## Sound Classes

Sounds which often occur in correspondence relations in genetically related languages can be clustered into classes (types). It is assumed “that phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’ ” (Dolgopolsky 1986: 35).

k

g

p

b

tʃ

dʒ

f

v

t

d

ʃ

ʒ

θ

ð

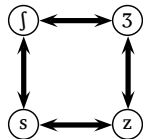
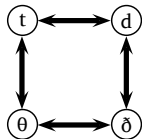
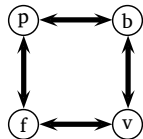
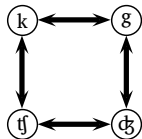
s

z

# Sound Classes

## Sound Classes

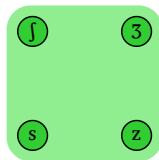
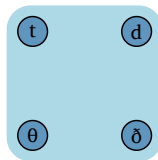
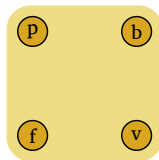
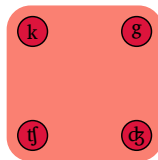
Sounds which often occur in correspondence relations in genetically related languages can be clustered into classes (types). It is assumed “that phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’ ” (Dolgopolsky 1986: 35).



# Sound Classes

## Sound Classes

Sounds which often occur in correspondence relations in genetically related languages can be clustered into classes (types). It is assumed “that phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’ ” (Dolgopolsky 1986: 35).



# Sound Classes

## Sound Classes

Sounds which often occur in correspondence relations in genetically related languages can be clustered into classes (types). It is assumed “that phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’ ” (Dolgopolsky 1986: 35).



## Scoring Functions for Sound Classes

- LingPy offers default scoring functions for three standard sound-class models (ASJP, SCA, DOLGO).
- The standard models vary regarding the roughness by which the continuum of sounds is split into discrete classes.
- The scoring functions are based on empirical data on sound correspondence frequencies (ASJP model, Brown et al. 2011), and on general theoretical models of the directionality and probability of sound change processes that are converted into non-directional similarity matrices (SCA, DOLGO, see List 2012 for details).

## Prosodic Strings

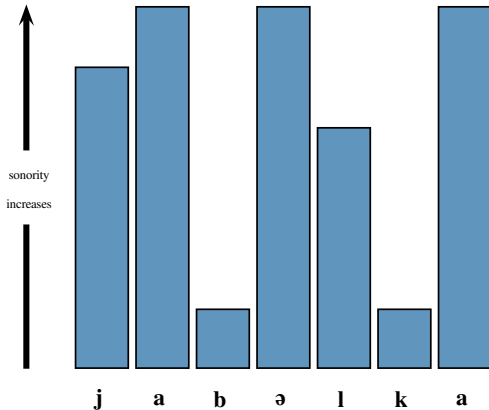
- Sound change occurs more frequently in prosodically **weak** positions of phonetic sequences (Geisler 1992).
- Given the **sonority profile** of a phonetic sequence, one can distinguish positions that differ regarding their **prosodic context**.
- Prosodic context can be modelled by representing a sequence by a **prosodic string**, indicating the different prosodic contexts of each segment.
- Based on the relative strength of all sites in a phonetic sequence, substitution scores and gap penalties can be modified when carrying out alignment analyses.
- Prosodic strings are an alternative to  $n$ -gram approaches, since they also handle context, their specific advantage being that they are more abstract and less data-dependent.



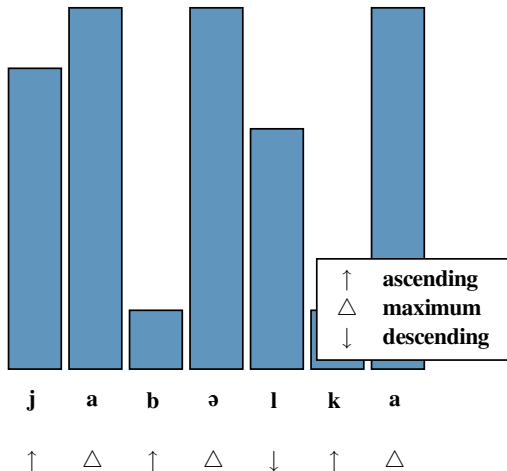
# Prosodic Strings

**j a b ə l k a**

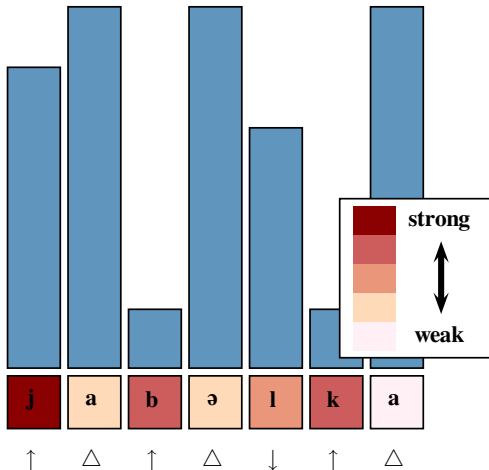
# Prosodic Strings



# Prosodic Strings



# Prosodic Strings



# Prosodic Strings

phonetic sequence	<b>j</b>	<b>a</b>	<b>b</b>	<b>ə</b>	<b>l</b>	<b>k</b>	<b>a</b>
SCA model	J	A	P	E	L	K	A
ASJP model	y	a	b	ɪ	l	k	a
DOLGO model	J	V	P	V	R	K	V
sonority profile	<b>6</b>	<b>7</b>	<b>1</b>	<b>7</b>	<b>5</b>	<b>1</b>	<b>7</b>
prosodic string	#	v	C	v	c	C	>
Relative Weight	<b>2.0</b>	<b>1.5</b>	<b>1.5</b>	<b>1.3</b>	<b>1.1</b>	<b>1.5</b>	<b>0.7</b>

## Secondary Alignment

- While secondary alignment was never an issue in computational biology, it is a desideratum in historical linguistics and dialectology. Secondary structures are especially important when
  - (1) aligning whole sentences, where the alignment of one word from one with two words from another sentence should be avoided,
  - (2) aligning language data for which morphological information is also available, or
  - (3) when aligning words from South-East-Asian tone languages which generally show a structure in which one syllable corresponds to one morpheme.

# Secondary Alignment

Primary Alignment						
<b>Haikou</b>	z	i	-	t	-	3
<b>Beijing</b>	z <sub>ɿ</sub>	ɿ	<sup>51</sup>	t <sup>h</sup>	ou	<sup>1</sup>

Secondary Alignment						
<b>Haikou</b>	z	i	t	<sup>3</sup>	-	-
<b>Beijing</b>	z <sub>ɿ</sub>	ɿ	-	<sup>51</sup>	t <sup>h</sup>	ou <sup>1</sup>

\*   \*  
v o l - d e m o r t  
v - l a d i m i r -  
v a l - d e m a r -  
\*   \*

Evaluation \*





# Evaluation Measures

- **PAS:** Perfect Alignment Score
- **CS:** Column Score
- **SPS:** Sum-of-Pairs Score

# Evaluation Measures

## Column-Score (CS)

$$CS = 100 \cdot 2 \cdot \frac{|C_t \cap C_r|}{|C_r| + |C_t|},$$

where  $C_t$  is the set of columns in the test alignment and  $C_r$  is the set of columns in the reference alignment (Rosenberg and Ogden 2009).

## Sum-of-Pairs Score (SPS)

$$SPS = 100 \cdot 2 \cdot \frac{|P_t \cap P_r|}{|P_r| + |P_t|},$$

where  $P_t$  is the set of all aligned residue pairs in the test alignment and  $P_r$  is the set of all aligned residue pairs in the reference alignment (ibd.).

# Gold Standard

- 1 089 manually aligned sequence pairs.
- Words taken from the Bai dialects (Wang 2006, Allen 2007) and Chinese dialects (Hou 2004).
- Both Bai and Chinese are tone languages.
- All data is available under  
<http://linguist.de/supp/secondary.zip>

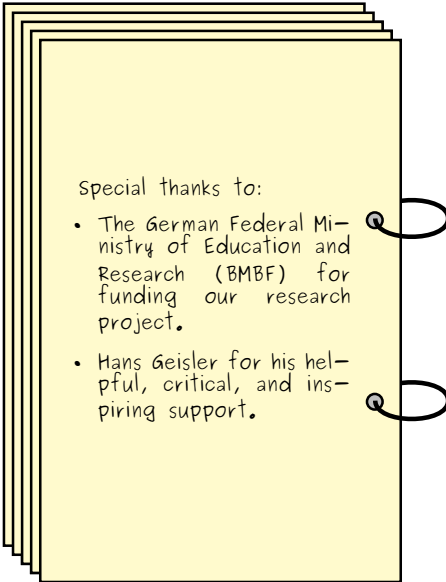
# Results

<b>Score</b>	<b>Primary</b>	<b>Secondary</b>
<b>PAS</b>	83.47	88.89
<b>CS</b>	88.54	92.70
<b>SPS</b>	92.78	95.52

## Concluding Remarks

- As can be seen from the results, the modified algorithm which is sensitive to secondary sequence structures shows a great improvement compared to the traditional algorithm which aligns sequences only with respect to their primary structure.
- The improvement is significant with  $p < 0.01$  using the Wilcoxon signed rank test as suggested by Notredame (2000).
- The algorithm for secondary alignment proves very useful for the alignment of tonal languages, yet it may also be employed for the analysis of other kinds of sequential data and, e.g., help to carry out phonetic alignment analyses of whole sentences.





Special thanks to:

- The German Federal Ministry of Education and Research (BMBF) for funding our research project.
- Hans Geisler for his helpful, critical, and inspiring support.

