

LexStat: Automatic Detection of Cognates in Multilingual Wordlists

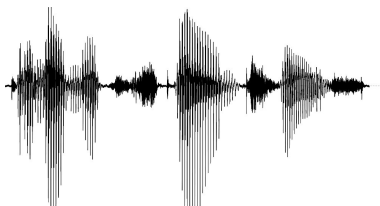
Johann-Mattis List*

*Institute for Romance Languages and Literature
Heinrich Heine University Düsseldorf

April 24, 2012

Structure of the Talk

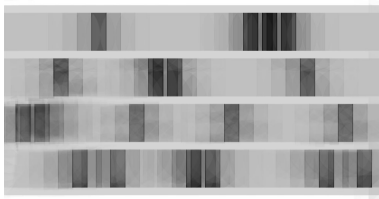
- 1 Keys to the Past
- 2 Identification of Cognates
- 3 LexStat
- 4 Evaluation



Do languages grow on trees?



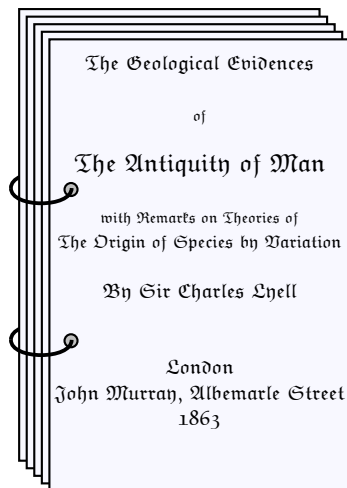
Keys to the Past



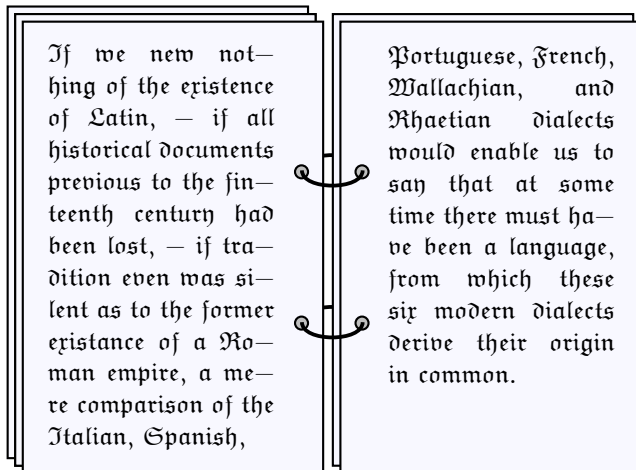
A A G T C T A G G T T A C C C G T A T

Charles Lyell on Languages

Charles Lyell on Languages



Charles Lyell on Languages



Uniformitarianism and Abduction

Uniformitarianism and Abduction

Uniformitarianism

Uniformitarianism and Abduction

Uniformitarianism

- “Universality of Change” – Change is independent of time and space

Uniformitarianism and Abduction

Uniformitarianism

- “Universality of Change” – Change is independent of time and space
- “Graduality of Change” – Change is neither abrupt nor chaotic

Uniformitarianism and Abduction

Uniformitarianism

- “Universality of Change” – Change is independent of time and space
- “Graduality of Change” – Change is neither abrupt nor chaotic
- “Uniformity of Change” – Change is not heterogeneous

Uniformitarianism and Abduction

Uniformitarianism

- “Universality of Change” – Change is independent of time and space
- “Graduality of Change” – Change is neither abrupt nor chaotic
- “Uniformity of Change” – Change is not heterogeneous

Abduction

Uniformitarianism and Abduction

Uniformitarianism

- “Universality of Change” – Change is independent of time and space
- “Graduality of Change” – Change is neither abrupt nor chaotic
- “Uniformity of Change” – Change is not heterogeneous

Abduction

- Present Events or Patterns
- + Known Laws
- => Abduction of Historical Facts

Uniformitarianism and Abduction

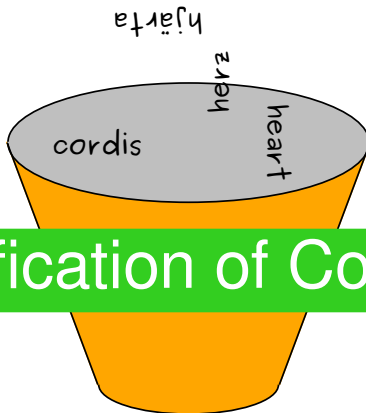
Uniformitarianism

- “Universality of Change” – Change is independent of time and space
- “Graduality of Change” – Change is neither abrupt nor chaotic
- “Uniformity of Change” – Change is not heterogeneous

Abduction

Present Events or Patterns
+ Known Laws
=> Abduction of Historical Facts

Similarities Between Languages
+ Language Change
=> Inference of Proto-Languages



Identification of Cognates

The Comparative Method

Basic Procedure

The Comparative Method

Basic Procedure

- Compile an initial list of putative cognate sets.

The Comparative Method

Basic Procedure

- Compile an initial list of putative cognate sets.
- Extract an initial list of putative sets of sound correspondences from the initial cognate list.

The Comparative Method

Basic Procedure

- 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

The Comparative Method

Basic Procedure

- 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

The Comparative Method

Basic Procedure

- 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.

The Comparative Method

Basic Procedure

- 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.

The Comparative Method

Language-Specific Similarity Measure

The Comparative Method

Language-Specific Similarity Measure

- Sequence similarity is determined on the basis of systematic sound correspondences as opposed to similarity based on surface resemblances of phonetic segments.

The Comparative Method

Language-Specific Similarity Measure

- 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 Comparative Method

Language-Specific Similarity Measure

- 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.

The Comparative Method

Language-Specific Similarity Measure

- 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 ʊ: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]

The Comparative Method

Language-Specific Similarity Measure

- 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”	[tɕ iɣ ³⁵]	Beijing [tɕ iou ²¹⁴]	[k ɐu ³⁵]
“today”	[tɕ iŋ ⁵⁵ tsɔ ²¹]	Beijing [tɕ iə ⁵⁵]	[k ɛm ⁵³ jet ²]
“rooster”	[kon ⁵⁵ tɕ i ²¹]	Beijing [kun ⁵⁵ tɕ i ⁵⁵]	[k ɛi ⁵⁵ kon ⁵⁵]

Automatic Approaches

Automatic Approaches

Alignment Analyses

Automatic Approaches

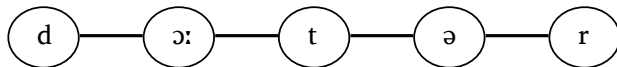
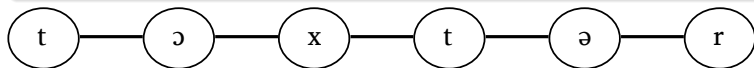
Alignment Analyses

In alignment analyses, sequences are arranged in a matrix in such a way that corresponding elements occur in the same column, while empty cells resulting from non-corresponding elements are filled with gap symbols.

Automatic Approaches

Alignment Analyses

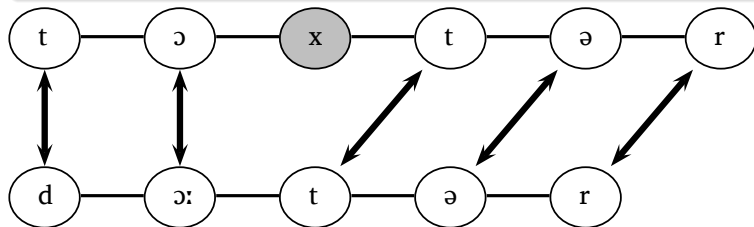
In alignment analyses, sequences are arranged in a matrix in such a way that corresponding elements occur in the same column, while empty cells resulting from non-corresponding elements are filled with gap symbols.



Automatic Approaches

Alignment Analyses

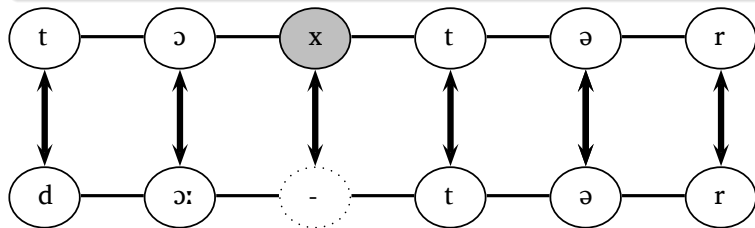
In alignment analyses, sequences are arranged in a matrix in such a way that corresponding elements occur in the same column, while empty cells resulting from non-corresponding elements are filled with gap symbols.



Automatic Approaches

Alignment Analyses

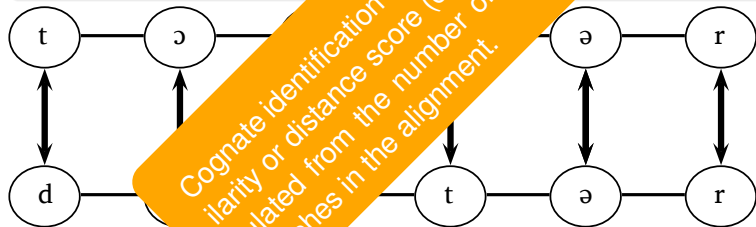
In alignment analyses, sequences are arranged in a matrix in such a way that corresponding elements occur in the same column, while empty cells resulting from non-corresponding elements are filled with gap symbols.



Automatic Approaches

Alignment Analyses

In alignment analyses, sequences are aligned in such a way that corresponding elements are in the same column, while empty cells represent gaps. Corresponding elements are filled with gaps.



Automatic Approaches

Sound Classes

Automatic Approaches

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).

Automatic Approaches

Sound Classes

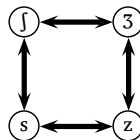
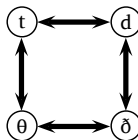
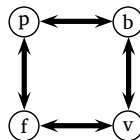
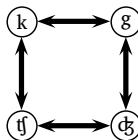
Sounds which often occur in correspondence relations in genetically related languages can be clustered into classes (types). It is assumed “that phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’” (Dolgopolsky 1986: 35).

Ⓚ	ⓖ	Ⓟ	Ⓟ
Ⓣ	Ⓣ	Ⓣ	Ⓟ
Ⓣ	Ⓣ	Ⓣ	Ⓣ
Ⓣ	Ⓣ	Ⓣ	Ⓣ

Automatic Approaches

Sound Classes

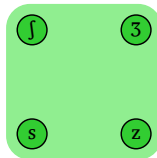
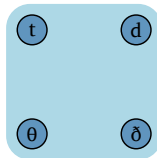
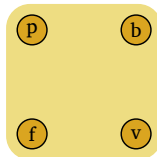
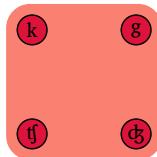
Sounds which often occur in correspondence relations in genetically related languages can be clustered into classes (types). It is assumed “that phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’” (Dolgopolsky 1986: 35).



Automatic Approaches

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).



Automatic Approaches

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****P****T****S**

Automatic Approaches

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 frequent than those between different 'types' (Dolgoplosky 1971).

Cognate identification is usually based on comparing the first two consonants of two words: If they match regarding their sound classes, the words are judged to be cognate, otherwise not.

P

T

S

Automatic Approaches

Sound-Class-Based Alignment (SCA)

Automatic Approaches

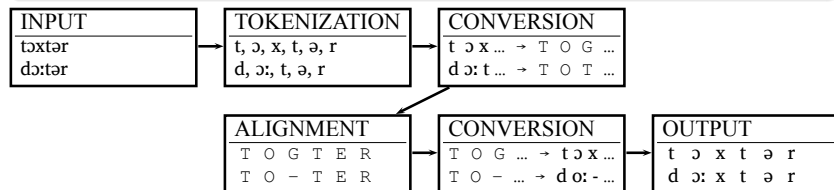
Sound-Class-Based Alignment (SCA)

Sound classes and alignment analyses can be easily combined by representing phonetic sequences internally as sound classes and comparing the sound classes with traditional alignment algorithms.

Automatic Approaches

Sound-Class-Based Alignment (SCA)

Sound classes and alignment analyses can be easily combined by representing phonetic sequences internally as sound classes and comparing the sound classes with traditional alignment algorithms.



Automatic Approaches

Sound-Class-Based Alignment (SCA)

Sound classes and alignment analysis are combined by representing phonetic sequences in terms of sound classes and comparing the sound class representations with alignment algorithms.



Cognate identification may be based on a certain threshold and distance scores derived from the similarity scores yielded by the alignment algorithm.

Traditional vs. Automatic Approaches

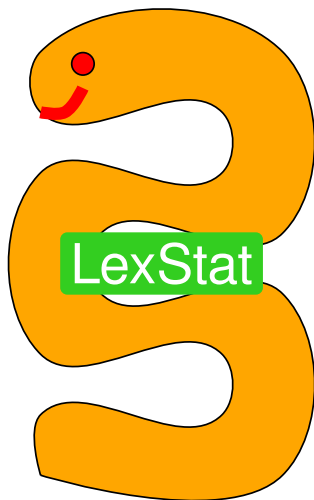
Traditional vs. Automatic Approaches

Similarity

Traditional vs. Automatic Approaches

Similarity

Almost all current automatic approaches are based on a language-independent similarity measure, while the comparative method applies a language-specific one. All automatic approaches will therefore yield the same scores for phenotypically identical sequences, regardless of the language systems they belong to.



Working Procedure

Working Procedure

Sequence Input

sequences are read from specifically formatted input files



Working Procedure

Sequence Input

sequences are read from specifically formatted input files

1 Sequence Conversion

sequences are converted to sound classes and prosodic profiles

Working Procedure

Sequence Input

sequences are read from specifically formatted input files

1 Sequence Conversion

sequences are converted to sound classes and prosodic profiles

2 Scoring-Scheme Creation

using a permutation method, language-specific scoring schemes are determined

Working Procedure

Sequence Input

sequences are read from specifically formatted input files

1 Sequence Conversion

sequences are converted to sound classes and prosodic profiles

2 Scoring-Scheme Creation

using a permutation method, language-specific scoring schemes are determined

3 Distance Calculation

based on the language-specific scoring-scheme, pairwise distances between sequences are calculated

Working Procedure

Sequence Input

sequences are read from specifically formatted input files

-
- 1 Sequence Conversion
sequences are converted to sound classes and prosodic profiles
 - 2 Scoring-Scheme Creation
using a permutation method, language-specific scoring schemes are determined
 - 3 Distance Calculation
based on the language-specific scoring-scheme, pairwise distances between sequences are calculated
 - 4 Sequence Clustering
sequences are clustered into cognate sets whose average distance is beyond a certain threshold

Working Procedure

Sequence Input

sequences are read from specifically formatted input files

1 Sequence Conversion

sequences are converted to sound classes and prosodic profiles

2 Scoring-Scheme Creation

using a permutation method, language-specific scoring schemes are determined

3 Distance Calculation

based on the language-specific scoring-scheme, pairwise distances between sequences are calculated

4 Sequence Clustering

sequences are clustered into cognate sets whose average distance is beyond a certain threshold

Sequence Output

information regarding sequence clustering is written to file using a specific format

Implementation

Implementation

- LexStat ist implemented as part of the LingPy Python library (see <http://linguist.de/lingpy>) for automatic tasks in historical linguistics.

Implementation

- LexStat is implemented as part of the LingPy Python library (see <http://linguist.de/lingpy>) for automatic tasks in historical linguistics.
- The current release of LingPy (lingpy-1.0) provides methods for pairwise and multiple sequence alignment (SCA), automatic cognate detection (LexStat), and plotting routines (see the online documentation for details).

Implementation

- LexStat is implemented as part of the LingPy Python library (see <http://linguist.de/lingpy>) for automatic tasks in historical linguistics.
- The current release of LingPy (lingpy-1.0) provides methods for pairwise and multiple sequence alignment (SCA), automatic cognate detection (LexStat), and plotting routines (see the online documentation for details).
- LexStat can be invoked from the Python shell or inside Python scripts (examples are given in the online documentation).

Input and Output

ID	Items	German	English	Swedish
1	hand	hant	hænd	hand
2	woman	fraʊ	wʊmən	kvina
3	know	kənən	nəʊ	çəna
3	know	visən	-	ve:ta
...	...	...	...	...

Input and Output

ID	Items	German	COG	English	COG	Swedish	COG
1	hand	hant	1	hænd	1	hand	1
2	woman	fraʊ	2	wʊmən	3	kvina	4
3	know	kenən	5	nəʊ	5	çena	5
3	know	visən	6	-	0	ve:ta	6
...	...	...	...	...	...	...	...

Input and Output

Basic Concept: <i>belly</i> (ID: 4)			
CogID	Language	Entry	Aligned Entry
6	Danish	ɔnʌliwʔ	--
7	German	baux	b au x
7	Dutch	bæyk	b æy k
7	Swedish	buk	b u k
7	Norwegian	bʉ:k	b ʉ: k
8	English	bɛlɪ	--
9	Swedish	ma:ge	m a: g e
9	Norwegian	mɑ:gə	m ɑ: g ə
9	Danish	mæ:və	m æ: v ə
10	Icelandic	kʰvɪ:ðʏr	--

Internal Representation of Sequences

Internal Representation of Sequences

Sound Classes and Prosodic Context

Internal Representation of Sequences

Sound Classes and Prosodic Context

- All sequences are internally represented as sound classes, the default model being the one proposed in List (forthcoming).

Internal Representation of Sequences

Sound Classes and Prosodic Context

- All sequences are internally represented as sound classes, the default model being the one proposed in List (forthcoming).
- All sequences are also represented by *prosodic strings* which indicate the prosodic environment (initial, ascending, maximum, descending, final) of each phonetic segment (List 2012).

Internal Representation of Sequences

Sound Classes and Prosodic Context

- All sequences are internally represented as sound classes, the default model being the one proposed in List (forthcoming).
- All sequences are also represented by *prosodic strings* which indicate the prosodic environment (initial, ascending, maximum, descending, final) of each phonetic segment (List 2012).
- The information regarding sound classes and prosodic context is combined, and each input sequence is further represented as a sequence of tuples, consisting of the sound class and the prosodic environment of the respective phonetic segment.

Scoring-Scheme Creation

Scoring-Scheme Creation

Attested Distribution

Scoring-Scheme Creation

Attested Distribution

- carry out global and pairwise alignment analyses of all sequence pairs occurring in the same semantic slot
- store all corresponding segments that occur in sequences whose distance is beyond a certain threshold

Scoring-Scheme Creation

Attested Distribution

- carry out global and pairwise alignment analyses of all sequence pairs occurring in the same semantic slot
- store all corresponding segments that occur in sequences whose distance is beyond a certain threshold

Creation of the Expected Distribution

Scoring-Scheme Creation

Attested Distribution

- carry out global and pairwise alignment analyses of all sequence pairs occurring in the same semantic slot
- store all corresponding segments that occur in sequences whose distance is beyond a certain threshold

Creation of the Expected Distribution

- shuffle the wordlists repeatedly and
 - carry out global and pairwise alignment analyses of all sequence pairs in the randomly shuffled wordlists
 - store all corresponding segments
- average the results

Scoring-Scheme Creation

Attested Distribution

- carry out global and pairwise alignment analyses of all sequence pairs occurring in the same semantic slot
- store all corresponding segments that occur in sequences whose distance is beyond a certain threshold

Creation of the Expected Distribution

- shuffle the wordlists repeatedly and
 - carry out global and pairwise alignment analyses of all sequence pairs in the randomly shuffled wordlists
 - store all corresponding segments
- average the results

Calculation of Similarity Scores

Scoring-Scheme Creation

Attested Distribution

- carry out global and pairwise alignment analyses of all sequence pairs occurring in the same semantic slot
- store all corresponding segments that occur in sequences whose distance is beyond a certain threshold

Creation of the Expected Distribution

- shuffle the wordlists repeatedly and
 - carry out global and pairwise alignment analyses of all sequence pairs in the randomly shuffled wordlists
 - store all corresponding segments
- average the results

Calculation of Similarity Scores

- Calculation of *log-odds scores* from the distributions.

Scoring-Scheme Creation

English	German	Att.	Exp.	Score
#[t,d]	#[t,d]	3.0	1.24	6.3
#[t,d]	#[ts]	3.0	0.38	6.0
#[t,d]	#[ʃ,s,z]	1.0	1.99	-1.5
#[θ,ð]	#[t,d]	7.0	0.72	6.3
#[θ,ð]	#[ts]	0.0	0.25	-1.5
#[θ,ð]	#[s,z]	0.0	1.33	0.5
[t,d]\$	[t,d]\$	21.0	8.86	6.3
[t,d]\$	[ts]\$	3.0	1.62	3.9
[t,d]\$	[ʃ,s]\$	6.0	5.30	1.5
[θ,ð]\$	[t,d]\$	4.0	1.14	4.8
[θ,ð]\$	[ts]\$	0.0	0.20	-1.5
[θ,ð]\$	[ʃ,s]\$	0.0	0.80	0.5

Scoring-Scheme Creation

English	German	Att.	Exp.	Score
#[t,d]	#[t,d]	3.0	1.24	6.3
#[t,d]	#[ts]	3.0	0.38	6.0
#[t,d]	#[ʃ,s,z]	1.0	1.99	-1.5
#[θ,ð]	#[t,d]	7.0	0.72	6.3
#[θ,ð]	#[ts]	0.0	0.25	-1.5
#[θ,ð]	#[s,z]	0.0	1.33	0.5
[t,d]\$	[t,d]\$	21.0	8.86	6.3
[t,d]\$	[ts]\$	3.0	1.62	3.9
[t,d]\$	[ʃ,s]\$	6.0	5.30	1.5
[θ,ð]\$	[t,d]\$	4.0	1.14	4.8
[θ,ð]\$	[ts]\$	0.0	0.20	-1.5
[θ,ð]\$	[ʃ,s]\$	0.0	0.80	0.5

Scoring-Scheme Creation

	Initial	Final
English	<i>town</i> [taʊn]	<i>hot</i> [hɒt]
German	<i>Zaun</i> [tsaun]	<i>heiß</i> [hais]
English	<i>thorn</i> [θɔ:n]	<i>mouth</i> [maʊθ]
German	<i>Dorn</i> [dɔrn]	<i>Mund</i> [mont]
English	<i>dale</i> [deɪl]	<i>head</i> [hɛd]
German	<i>Tal</i> [ta:l]	<i>Hut</i> [hʊt]

Sequence Clustering

	Ger.	Eng.	Dan.	Swe.	Dut.	Nor.
Ger. [frau]	0.00	0.95	0.81	0.70	0.34	1.00
Eng. [wʊmən]	0.95	0.00	0.78	0.90	0.80	0.80
Dan. [kvenə]	0.81	0.78	0.00	0.17	0.96	0.13
Swe. [kvin:a]	0.70	0.90	0.17	0.00	0.86	0.10
Dut. [vrəʊv]	0.34	0.80	0.96	0.86	0.00	0.89
Nor. [kvinə]	1.00	0.80	0.13	0.10	0.89	0.00

Sequence Clustering

	Ger.	Eng.	Dan.	Swe.	Dut.	Nor.
Ger. [frau]	0.00	0.95	0.81	0.70	0.34	1.00
Eng. [wʊmən]	0.95	0.00	0.78	0.90	0.80	0.80
Dan. [kvenə]	0.81	0.78	0.00	0.17	0.96	0.13
Swe. [kvin:a]	0.70	0.90	0.17	0.00	0.86	0.10
Dut. [vrəʊv]	0.34	0.80	0.96	0.86	0.00	0.89
Nor. [kvinə]	1.00	0.80	0.13	0.10	0.89	0.00
Clusters	1	2	3	3	1	3

* *

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 *

*



Gold Standard

Gold Standard

File	Family	Lng.	Itm.	Entr.	Source
GER	Germanic	7	110	814	Starostin (2008)
ROM	Romance	5	110	589	Starostin (2008)
SLV	Slavic	4	110	454	Starostin (2008)
PIE	Indo-Eur.	18	110	2057	Starostin (2008)
OUG	Uralic	21	110	2055	Starostin (2008)
BAI	Bai	9	110	1028	Wang (2006)
SIN	Sinitic	9	180	1614	Hóu (2004)
KSL	varia	8	200	1600	Kessler (2001)
JAP	Japonic	10	200	1986	Shirō (1973)

Evaluation Measures

Evaluation Measures

Set Comparison

Evaluation Measures

Set Comparison

Precision, Recall, and F-Score are calculated by comparing the cognate sets proposed by the method with the cognate sets in the gold standard (see Bergsma & Kondrak 2007).

Evaluation Measures

Set Comparison

Precision, Recall, and F-Score are calculated by comparing the cognate sets proposed by the method with the cognate sets in the gold standard (see Bergsma & Kondrak 2007).

Pair Comparison

Evaluation Measures

Set Comparison

Precision, Recall, and F-Score are calculated by comparing the cognate sets proposed by the method with the cognate sets in the gold standard (see Bergsma & Kondrak 2007).

Pair Comparison

Pair comparison is based on a pairwise comparison of all decisions present in testset and goldstandard.

Tests

Tests

- Sound Classes – matching sound classes without alignment (based on Turchin et al. 2010)

Tests

- Sound Classes – matching sound classes without alignment (based on Turchin et al. 2010)
- Simple Alignment – normalized edit-distance (Levenshtein 1966)

Tests

- Sound Classes – matching sound classes without alignment (based on Turchin et al. 2010)
- Simple Alignment – normalized edit-distance (Levenshtein 1966)
- SCA – language-independent distance scores derived from sound-class-based alignment analyses (List 2012)

Tests

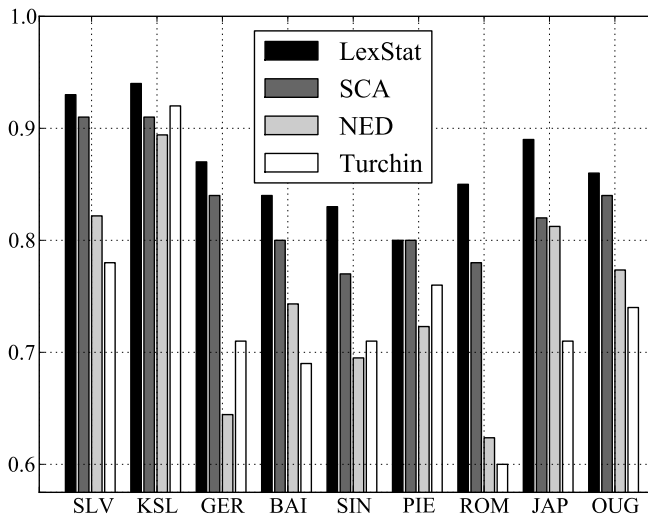
- Sound Classes – matching sound classes without alignment (based on Turchin et al. 2010)
- Simple Alignment – normalized edit-distance (Levenshtein 1966)
- SCA – language-independent distance scores derived from sound-class-based alignment analyses (List 2012)
- LexStat – language-specific distance scores

General Results

General Results

Score	LexStat	SCA	Simple Alm.	Sound Cl.
Identical Pairs	0.85	0.82	0.76	0.74
Precision	0.59	0.51	0.39	0.39
Recall	0.68	0.57	0.47	0.55
F-Score	0.63	0.55	0.42	0.46

General Results



Specific Results

Specific Results

- Pairwise decisions were extracted from the KSL dataset and compared with the Gold Standard.

Specific Results

- Pairwise decisions were extracted from the KSL dataset and compared with the Gold Standard.
- 72 borrowings were explicitly marked along with their source by Kessler (2001).

Specific Results

- Pairwise decisions were extracted from the KSL dataset and compared with the Gold Standard.
- 72 borrowings were explicitly marked along with their source by Kessler (2001).
- 83 chance resemblances were determined automatically by taking non-cognate word pairs with an NED score less than 0.6.

Specific Results

- Pairwise decisions were extracted from the KSL dataset and compared with the Gold Standard.
- 72 borrowings were explicitly marked along with their source by Kessler (2001).
- 83 chance resemblances were determined automatically by taking non-cognate word pairs with an NED score less than 0.6.

	LexStat	SCA	Simple Alm.	Sound Cl.
Borrowings	50%	61%	49%	53%
Chance Resemblances	17%	42%	89%	31%



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.
- James Kilbury for all the time he spent on helping me to refine the manuscript.

