Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists

Gerhard Jäger[1], Johann-Mattis List[2] & Pavel Sofroniev[1]

[1]Tübingen University & [2]MPI Jena

Valencia, EACL 2017

April 7, 2017

WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

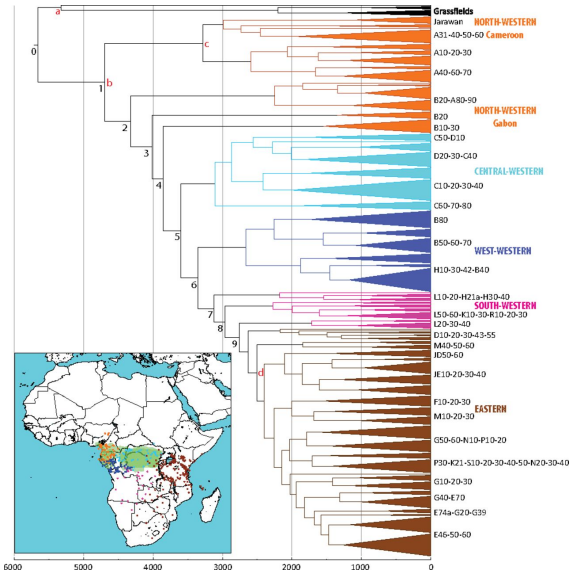
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

DFG

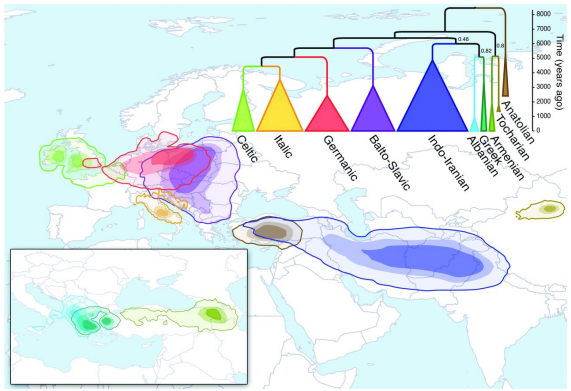
erc
European Research Council
Established by the European Commission

Computational historical linguistics

- massive progress within past 15 years
- automated language classification
- inferring time depth and homeland of language families
- automatic reconstruction of proto-languages
- discovery of statistical patterns in language change
- ...



(Grollemund et al, 2015)



(Bouckaert et al, 2012)

Computational historical linguistics

- most work depends on manually coded cognate judgments on Swadesh lists
  - labor intensive
  - subjective
  - not fully replicable
  - induces bias in favor of well-studied language families

Cognate-coded word lists

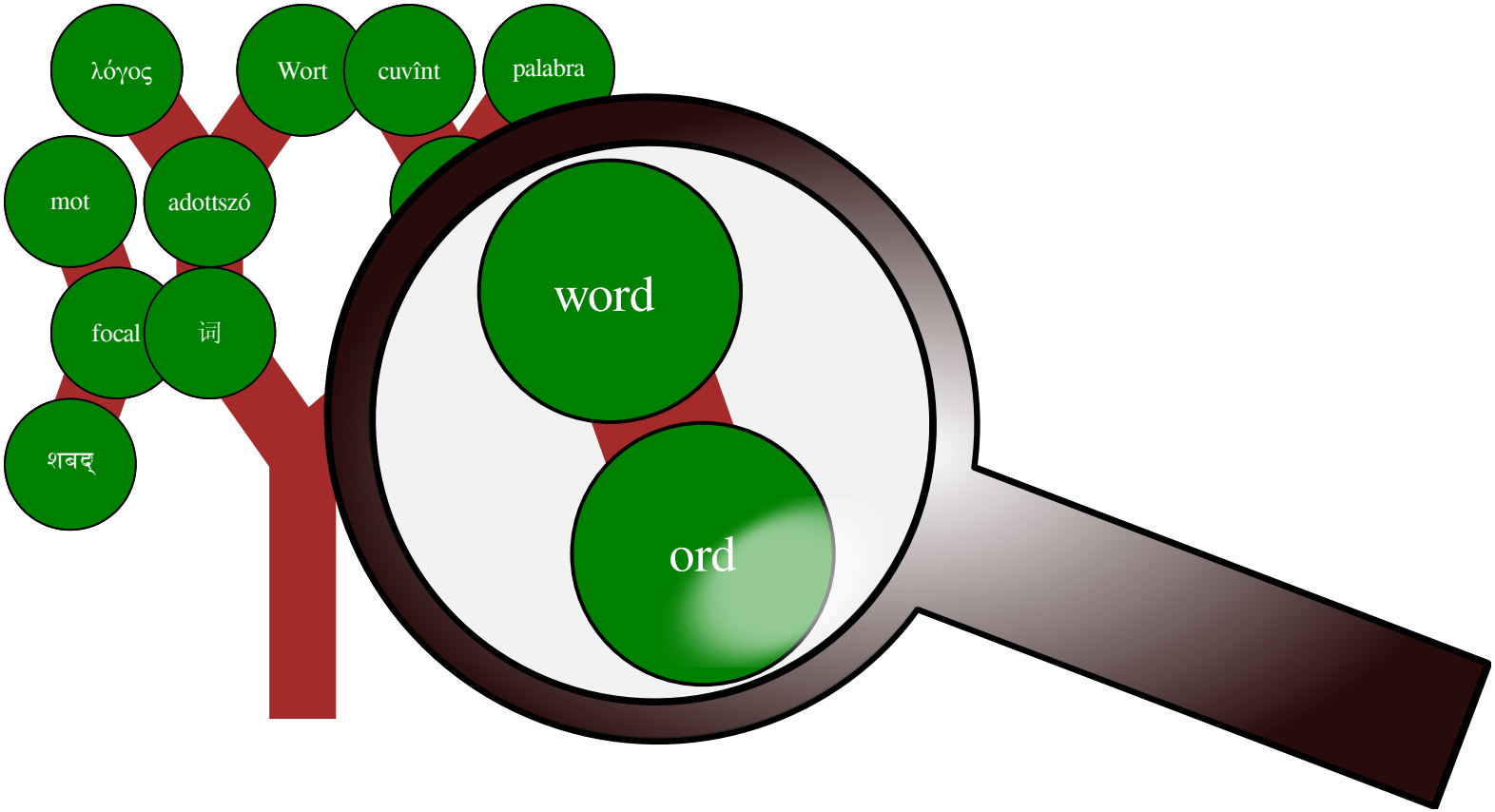
- typical data structure in CHL

language	*iso*	*gloss*	*gloss_id*	*transcr.*	*cogn._class*
ELFDALIAN	qov	woman	962	ˈkɛ̀lɪŋg	woman:Ag
DUTCH	nld	woman	962	vrɑu	woman:B
GERMAN	deu	woman	962	fraŭ	woman:B
DANISH	dan	woman	962	ˈg̊ʰvenə	woman:D
DANISH_FJOLDE		woman	962	kvinʲ	woman:D
GUTNISH_LAU		woman	962	ˈkvɪnːˌfolk	woman:D
LATIN	lat	woman	962	ˈmulier	woman:E
LATIN	lat	woman	962	ˈfeːmina	woman:G
ENGLISH	eng	woman	962	wʊmən	woman:H
GERMAN	deu	woman	962	vaĭp	woman:H
DANISH	dan	woman	962	ˈd̥ɛːmə	woman:K

- goal of this talk:

  How to automatically infer cognate classification.

Previous Work



Previous Work

ID	Taxa	Word	Gloss	GlossID	IPA	...
...	...	...	...	...	...	...
21	German	Frau	woman	20	frau	...
22	Dutch	vrouw	woman	20	vrɑu	...
23	English	woman	woman	20	wʊmən	...
24	Danish	kvinde	woman	20	kvenə	...
25	Swedish	kvinna	woman	20	kviːna	...
26	Norwegian	kvine	woman	20	kʉinə	...
...	...	...	...	...	...	...

Previous Work

ID	Taxa	Word	Gloss	GlossID	IPA	CogID
...	...	...	...	...	...	...
21	German	Frau	woman	20	frau	1
22	Dutch	vrouw	woman	20	vrɑu	1
23	English	woman	woman	20	wʊmən	2
24	Danish	kvinde	woman	20	kvenə	3
25	Swedish	kvinna	woman	20	kviːna	3
26	Norwegian	kvine	woman	20	kʊinə	3
...	...	...	...	...	...	...

Previous Work

ID	Taxa	Word	Gloss	GlossID	IPA	CogID
...	...	...	...	...	...	...
21	German	Frau	woman	20	frau	1
22	Dutch	vrouw	woman	20	vrɑu	1
23	English	woman	woman	20	wʊmən	2
24	Danish	kvinde	woman	20	kvenə	3
25	Swedish	kvinna	woman	20	kviːna	3
26	Norwegian	kvine	woman	20	kʊinə	3
...	...	...	...	...	...	...

Previous Work: Sound Classes

Sound Classes

Previous Work: 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).

Previous Work: Sound Classes

Sound Classes

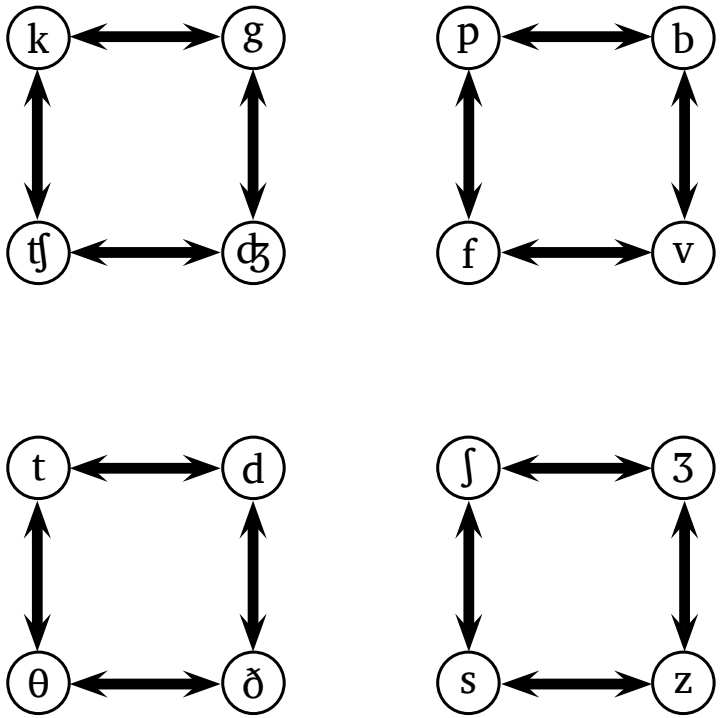
Sounds which often occur in correspondence relations in genetically related languages can be clustered into classes (types). It is assumed "that phonetic correspondences inside a 'type' are more regular than those between different 'types'" (Dolgopolsky 1986: 35).

k	g	p	b
tʃ	dʒ	f	v
t	d	ʃ	ʒ
θ	ð	s	z

Previous Work: Sound Classes

Sound Classes

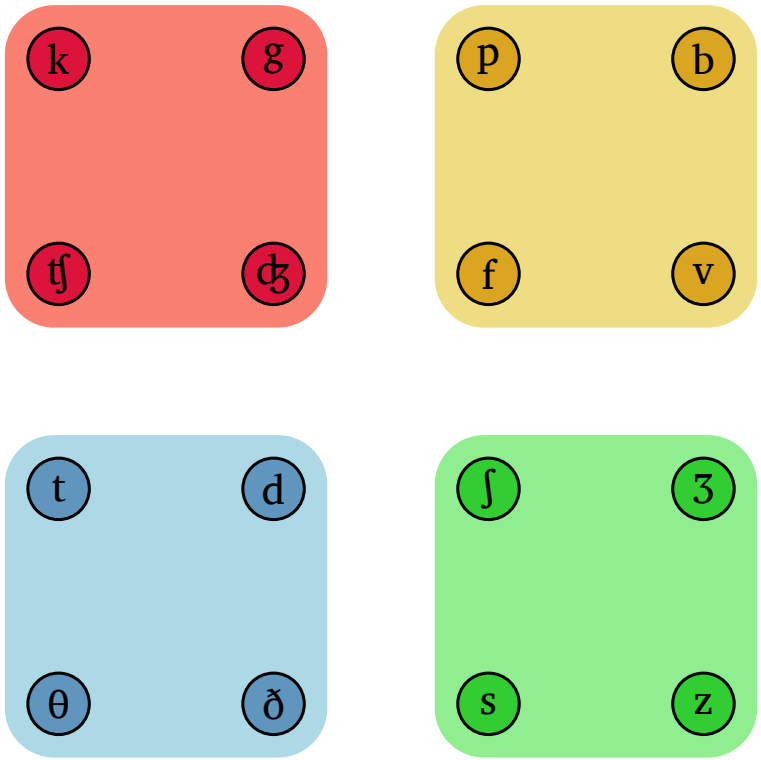
Sounds which often occur in correspondence relations in genetically related languages can be clustered into classes (types). It is assumed "that phonetic correspondences inside a 'type' are more regular than those between different 'types'" (Dolgopolsky 1986: 35).



Previous Work: Sound Classes

Sound Classes

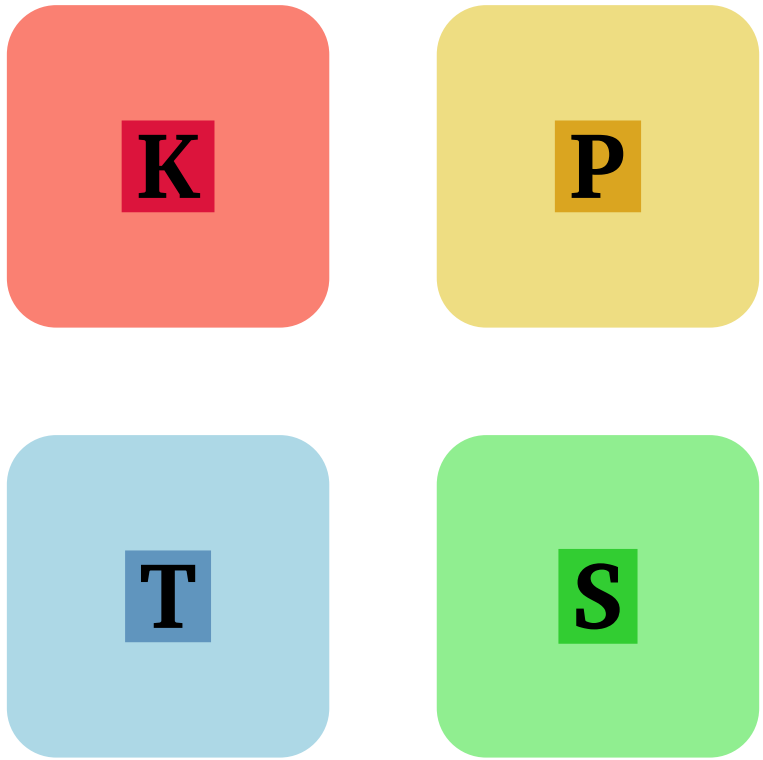
Sounds which often occur in correspondence relations in genetically related languages can be clustered into classes (types). It is assumed "that phonetic correspondences inside a 'type' are more regular than those between different 'types'" (Dolgopolsky 1986: 35).



Previous Work: 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).



Previous Work: Sound Classes

Sound Classes

Sounds which often occu
correspondence r
geneticall
cluster
assumed
correspor
more regu
different 'ty
35).

Ky 1986:

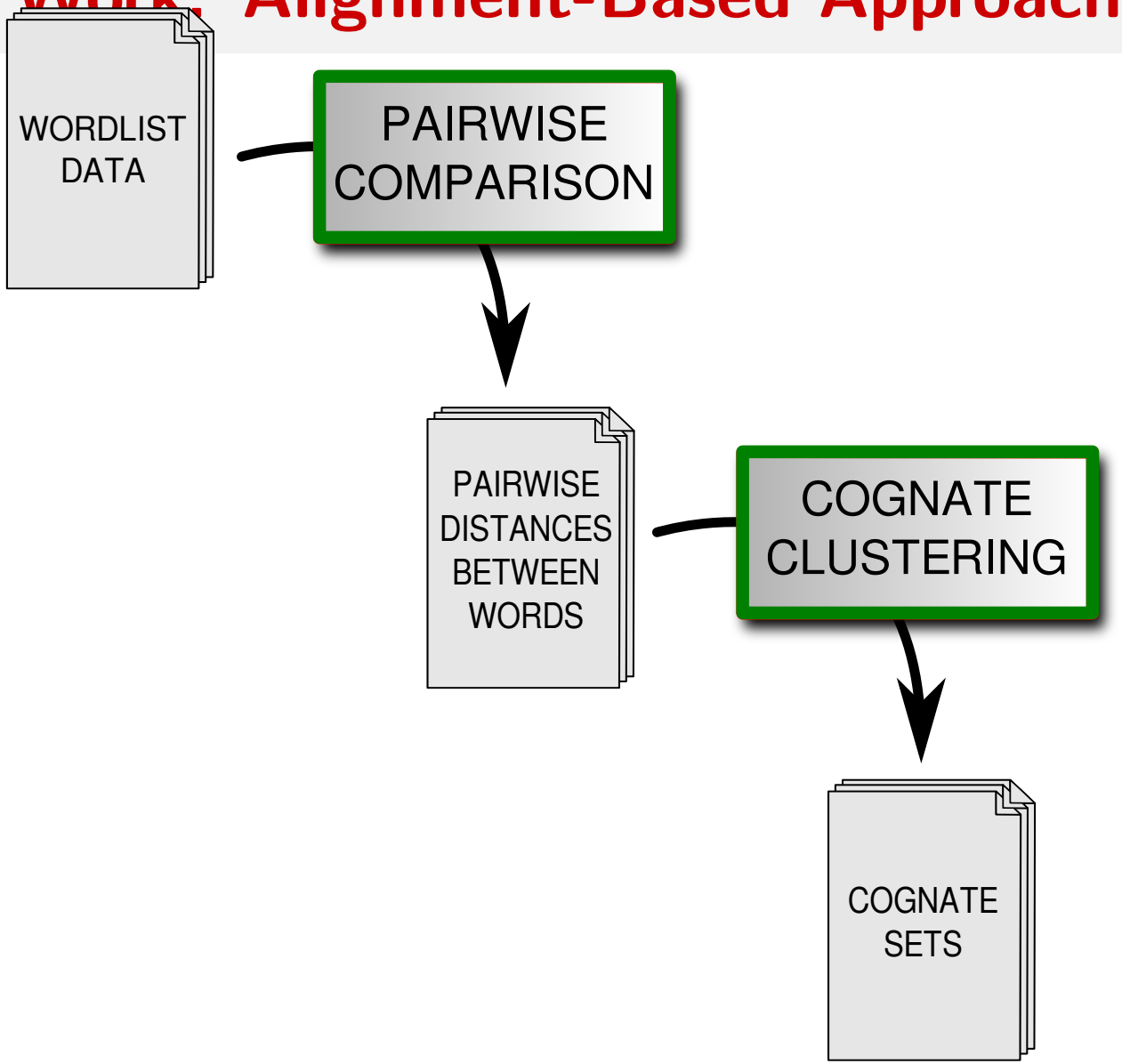
Cognate identification according to the Consonant-Class Matching (CCM) approach is usually based on comparing the first two consonants of two words: If they match regarding their sound classes, the words are judged to be cognate, otherwise not.

P

T

S

Previous Work: Alignment-Based Approaches



Previous Work: Alignment-Based Approaches

ID	Taxa	Word	Gloss	GlossID	IPA
...	...	...	...	...	...
21	German	Frau	woman	20	frau
22	Dutch	vrouw	woman	20	vrɑu
23	English	woman	woman	20	wʊmən
24	Danish	kvinde	woman	20	kvenə
25	Swedish	kvinna	woman	20	kvi:na
26	Norwegian	kvine	woman	20	kvinə
...	...	...	...	...	...

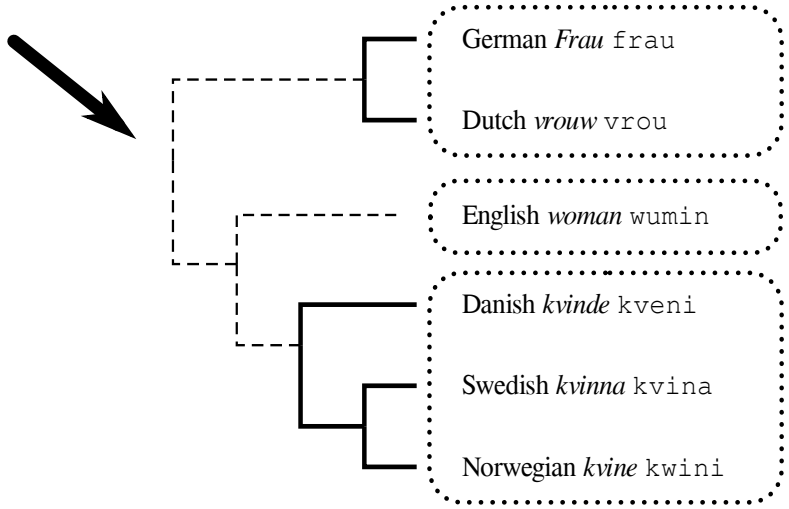
Previous Work: Alignment-Based Approaches

ID	Taxa	Word	Gloss	GlossID	IPA
...	...	...	...	...	...
21	German	Frau	woman	20	frau
22	Dutch	vrouw	woman	20	vrɑu
23	English	woman	woman	20	wʊ̈mən
24	Danish	kvinde	woman	20	kvenə
25	Swedish	kvinna	woman	20	kvi:na
26	Norwegian	kvine	woman	20	kvinə
...	...	...	...	...	...

	Swedish *kvinna*	English *woman*	Danish *kvinde*	Norwegian *kvine*	Dutch *vrouw*	German *Frau*
Swedish kvina	0.00	0.69	0.07	0.12	0.71	0.78
English wumin	0.69	0.00	0.66	0.57	0.68	0.87
Danish kveni	0.07	0.66	0.00	0.08	0.67	0.71
Norwegian kwini	0.12	0.57	0.08	0.00	0.75	0.74
Dutch frou	0.71	0.68	0.67	0.75	0.00	0.17
German frau	0.78	0.87	0.71	0.74	0.17	0.00

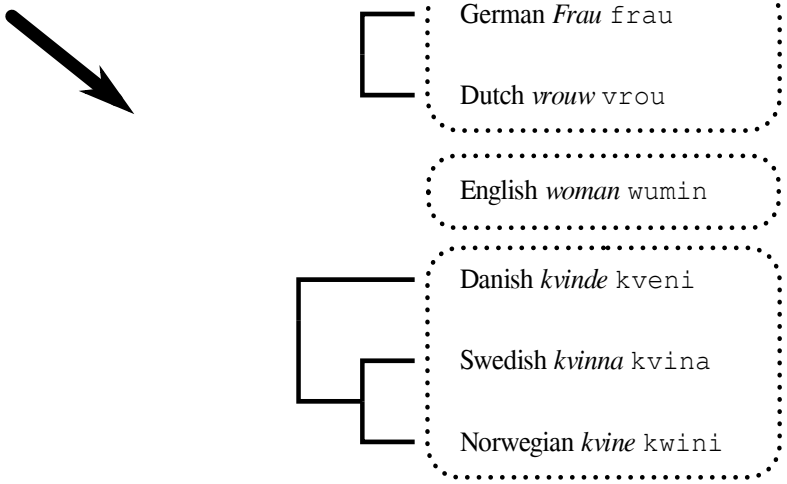
Previous Work: Alignment-Based Approaches

	Swedish *kvinna*	English *woman*	Danish *kvinde*	Norwegian *kvine*	Dutch *vrouw*	German *Frau*
Swedish `kvina`	0.00	0.69	0.07	0.12	0.71	0.78
English `wumin`	0.69	0.00	0.66	0.57	0.68	0.87
Danish `kveni`	0.07	0.66	0.00	0.08	0.67	0.71
Norwegian `kwini`	0.12	0.57	0.08	0.00	0.75	0.74
Dutch `frou`	0.71	0.68	0.67	0.75	0.00	0.17
German `frau`	0.78	0.87	0.71	0.74	0.17	0.00



Previous Work: Alignment-Based Approaches

	Swedish *kvinna*	English *woman*	Danish *kvinde*	Norwegian *kvine*	Dutch *vrouw*	German *Frau*
Swedish `kvina`	0.00	0.69	0.07	0.12	0.71	0.78
English `wumin`	0.69	0.00	0.66	0.57	0.68	0.87
Danish `kveni`	0.07	0.66	0.00	0.08	0.67	0.71
Norwegian `kwini`	0.12	0.57	0.08	0.00	0.75	0.74
Dutch `frou`	0.71	0.68	0.67	0.75	0.00	0.17
German `frau`	0.78	0.87	0.71	0.74	0.17	0.00



Materials

Goldstandard data

- various collections of phonetically transcribed and cognate-coded Swadesh lists from the literature
- pre-processing
  - correction of errata
  - replacement of non-IPA symbols by IPA counterparts
  - removal of non-IPA symbols conveying meta-information
  - removal of morphological markers
  - mapping of concepts to standardized concept sets from Concepticon (List et al., 2016)

Dataset	Words	Conc.	Lang.	Families	Cog.	Div.
ABVD (Greenhill et al. 2008)	12414	210	100	Austronesian	3558	0.27
Afrasian (Militarev 2000)	790	40	21	Afro-Asiatic	355	0.42
Bai (Wang 2006)	1028	110	9	Sino-Tibetan	285	0.19
Chinese (Hóu 2004)	2789	140	15	Sino-Tibetan	1189	0.40
Chinese (Běijīng Dàxué 1964)	3632	179	18	Sino-Tibetan	1225	0.30
Huon (McElhanon 1967)	1176	84	14	Trans-New Guinea	537	0.41
IELex (Dunn 2012)	11479	208	52	Indo-European	2459	0.20
Japanese (Hattori 1973)	1983	199	10	Japonic	456	0.15
Kadai (Peiros 1998)	400	40	12	Tai-Kadai	103	0.17
Kamasau (Sanders 1980)	271	36	8	Torricelli	60	0.10
Lolo-Burmese (Peiros 1998)	570	40	15	Sino-Tibetan	101	0.12
Central Asian (Manni et al. 2016)	15903	183	88	Altaic (Turkic), Indo-European	895	0.05
Mayan (Brown 2008)	2841	100	30	Mayan	844	0.27
Miao-Yao (Peiros 1998)	208	36	6	Hmong-Mien	70	0.20
Mixe-Zoque (Cysouw et al. 2006)	961	100	10	Mixe-Zoque	300	0.23
Mon-Khmer (Peiros 1998)	1424	100	16	Austroasiatic	719	0.47
ObUgrian (Zhivlov 2011)	2006	110	21	Uralic	229	0.06
Tujia (Starostin 2013)	498	107	5	Sino-Tibetan	164	0.15

Materials

Goldstandard data

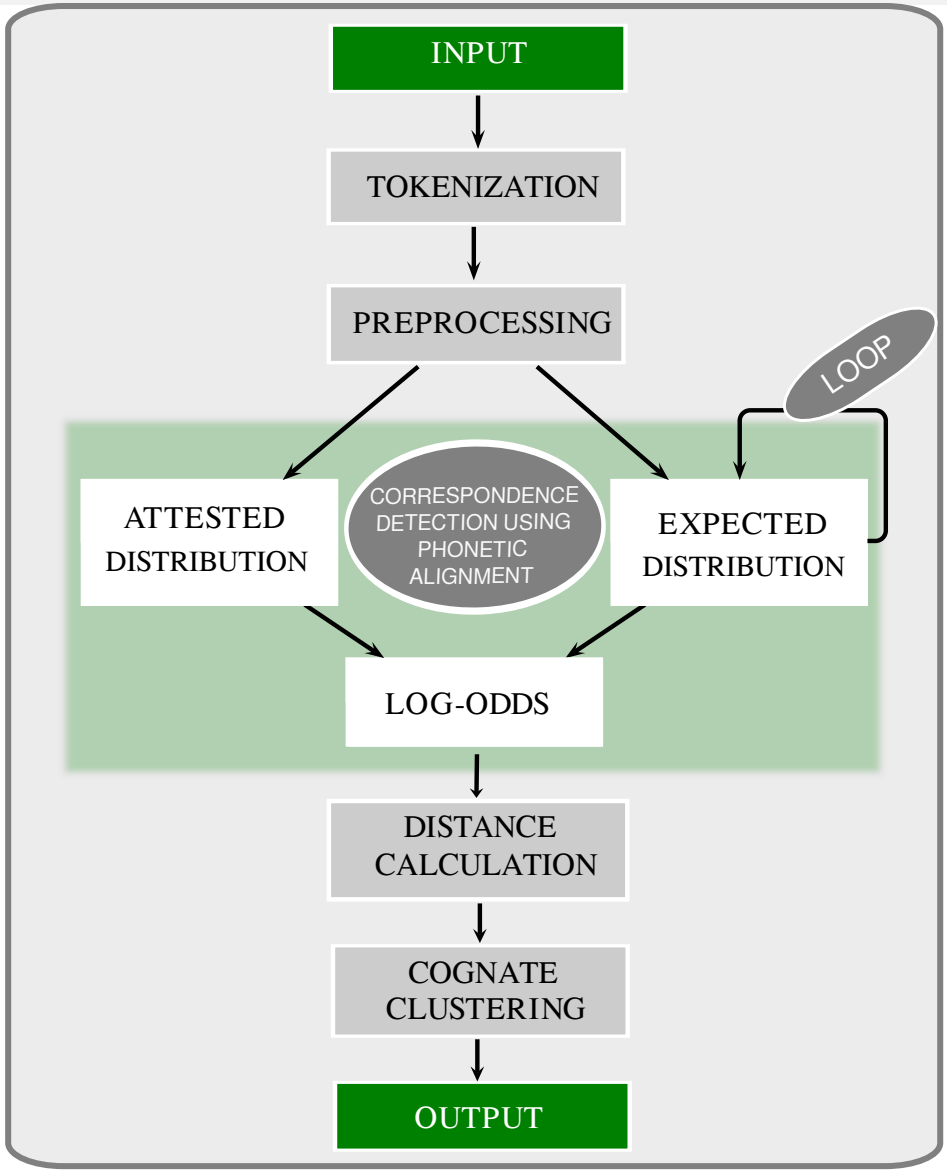
- three largest datasets used for testing
- all other datasets used for training

Dataset	Words	Conc.	Lang.	Families	Cog.	Div.
ABVD (Greenhill et al. 2008)	12414	210	100	Austronesian	3558	0.27
Afrasian (Militarev 2000)	790	40	21	Afro-Asiatic	355	0.42
Bai (Wang 2006)	1028	110	9	Sino-Tibetan	285	0.19
Chinese (Hóu 2004)	2789	140	15	Sino-Tibetan	1189	0.40
Chinese (Běijīng Dàxué 1964)	3632	179	18	Sino-Tibetan	1225	0.30
Huon (McElhanon 1967)	1176	84	14	Trans-New Guinea	537	0.41
IELex (Dunn 2012)	11479	208	52	Indo-European	2459	0.20
Japanese (Hattori 1973)	1983	199	10	Japonic	456	0.15
Kadai (Peiros 1998)	400	40	12	Tai-Kadai	103	0.17
Kamasau (Sanders 1980)	271	36	8	Torricelli	60	0.10
Lolo-Burmese (Peiros 1998)	570	40	15	Sino-Tibetan	101	0.12
Central Asian (Manni et al. 2016)	15903	183	88	Altaic (Turkic), Indo-European	895	0.05
Mayan (Brown 2008)	2841	100	30	Mayan	844	0.27
Miao-Yao (Peiros 1998)	208	36	6	Hmong-Mien	70	0.20
Mixe-Zoque (Cysouw et al. 2006)	961	100	10	Mixe-Zoque	300	0.23
Mon-Khmer (Peiros 1998)	1424	100	16	Austroasiatic	719	0.47
ObUgrian (Zhivlov 2011)	2006	110	21	Uralic	229	0.06
Tujia (Starostin 2013)	498	107	5	Sino-Tibetan	164	0.15

LexStat

- algorithm first propose in List (2012) and then further enhanced in List (2014), List et al. (2016) and List et al. (2017)
- the algorithm is generally based on the alignment-based workflow for cognate detections
- implemented as part of LingPy (lingpy.org, List and Forkel 2016)
- improvements include
  - scoring functions for alignments are computed *individually* for each language pair, modeling regular sound correspondences in classical linguistics
  - scores for both global and local alignment analyses are combined and agglomerated
  - alignment algorithm is sensitive for morpheme boundaries if they are annotated (*secondary alignment*, List 2014)
  - sequences are represented as *multi-tiered structures* which allows to handle *prosodic context*
  - agglomerative clustering procedure has been replaced by a community detection algorithm (Infomap, Rosvall and Bergstrom 2007)

LexStat



LexStat Algorithm (List 2014)

LexStat: Prosodic Strings

Prosodic Strings

Sound change occurs more frequently in weak phonotactic positions (Geisler 1992). Based on the sonority profile of a sound sequence, we can determine positions which differ with respect to their prosodic environment. Prosodic context can be modeled as prosodic string which distinguishes different contexts, and added as second tier to the sequence.

LexStat: Prosodic Strings

Prosodic Strings

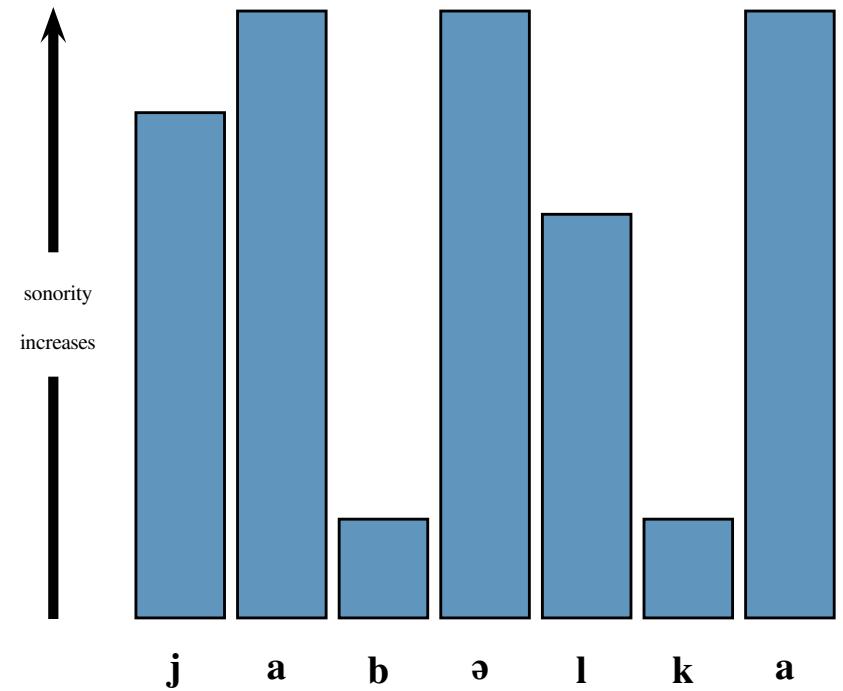
Sound change occurs more frequently in weak phonotactic positions (Geisler 1992). Based on the sonority profile of a sound sequence, we can determine positions which differ with respect to their prosodic environment. Prosodic context can be modeled as prosodic string which distinguishes different contexts, and added as second tier to the sequence.

j a b ə l k a

LexStat: Prosodic Strings

Prosodic Strings

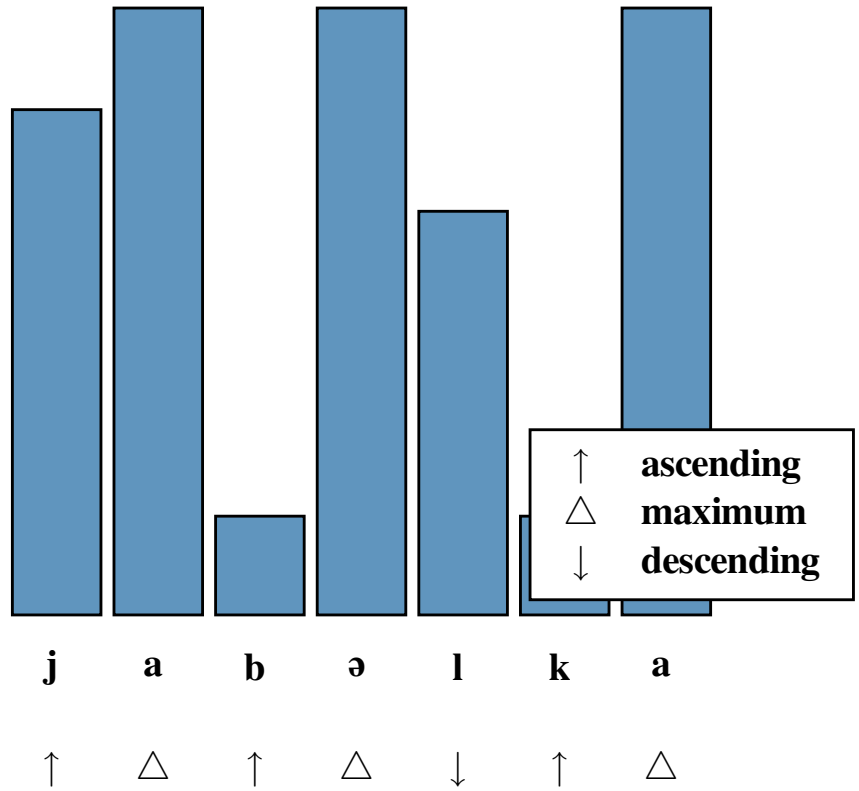
Sound change occurs more frequently in weak phonotactic positions (Geisler 1992). Based on the sonority profile of a sound sequence, we can determine positions which differ with respect to their prosodic environment. Prosodic context can be modeled as prosodic string which distinguishes different contexts, and added as second tier to the sequence.



LexStat: Prosodic Strings

Prosodic Strings

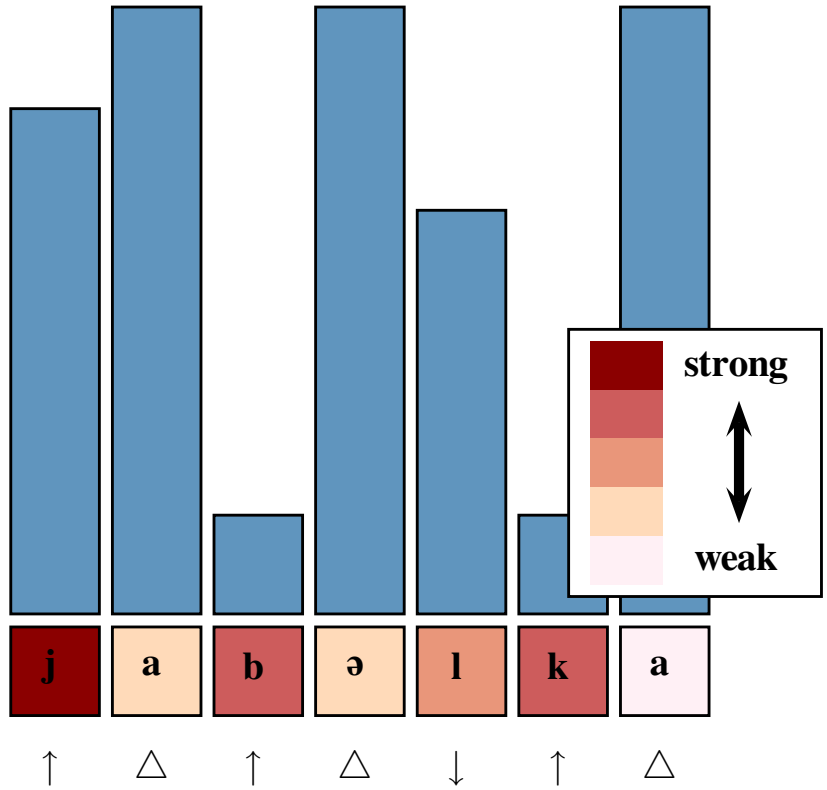
Sound change occurs more frequently in weak phonotactic positions (Geisler 1992). Based on the sonority profile of a sound sequence, we can determine positions which differ with respect to their prosodic environment. Prosodic context can be modeled as prosodic string which distinguishes different contexts, and added as second tier to the sequence.



LexStat: Prosodic Strings

Prosodic Strings

Sound change occurs more frequently in weak phonotactic positions (Geisler 1992). Based on the sonority profile of a sound sequence, we can determine positions which differ with respect to their prosodic environment. Prosodic context can be modeled as prosodic string which distinguishes different contexts, and added as second tier to the sequence.



LexStat: Prosodic Strings

Prosodic Strings

Sound change occurs more frequently in weak phonotactic positions (Geisler 1992). Based on the sonority profile of a sound sequence, we can determine positions which differ with respect to their prosodic environment. Prosodic context can be modeled as prosodic string which distinguishes different contexts, and added as second tier to the sequence.

j	a	b	ə	l	k	a
#	V	C	V	C	C	>

LexStat: Language-specific scoring schemes

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

LexStat: Language-specific scoring schemes

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

LexStat: Language-specific scoring schemes

	Initial	Final
English	*town* [taʊn]	*hot* [hɔt]
German	*Zaun* [tsaun]	*heiß* [haɪs]
English	*thorn* [θɔːn]	*mouth* [maʊθ]
German	*Dorn* [dɔrn]	*Mund* [mʊnt]
English	*dale* [deɪl]	*head* [hɛd]
German	*Tal* [taːl]	*Hut* [huːt]

LexStat: Cognate Set Partitioning

A

		GREEK	GERMAN	ENGLISH	RUSSIAN	POLISH
GREEK	çeri	0.00	0.72	0.69	0.73	0.77
GERMAN	hant	0.72	0.00	0.03	0.91	0.70
ENGLISH	hænd	0.69	0.03	0.00	0.91	0.68
RUSSIAN	ruka	0.72	0.91	0.91	0.00	0.20
POLISH	rɛ̃ŋka	0.77	0.70	0.68	0.20	0.00

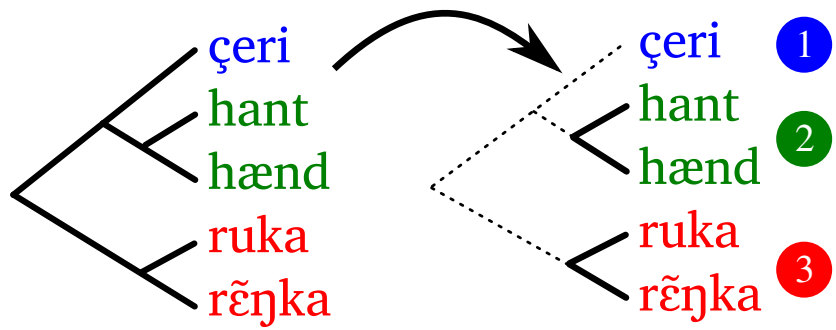
List et al. 2017

LexStat: Cognate Set Partitioning

A

		GREEK	GERMAN	ENGLISH	RUSSIAN	POLISH
GREEK	çeri	0.00	0.72	0.69	0.73	0.77
GERMAN	hant	0.72	0.00	0.03	0.91	0.70
ENGLISH	hænd	0.69	0.03	0.00	0.91	0.68
RUSSIAN	ruka	0.72	0.91	0.91	0.00	0.20
POLISH	rɛ̃ŋka	0.77	0.70	0.68	0.20	0.00

B



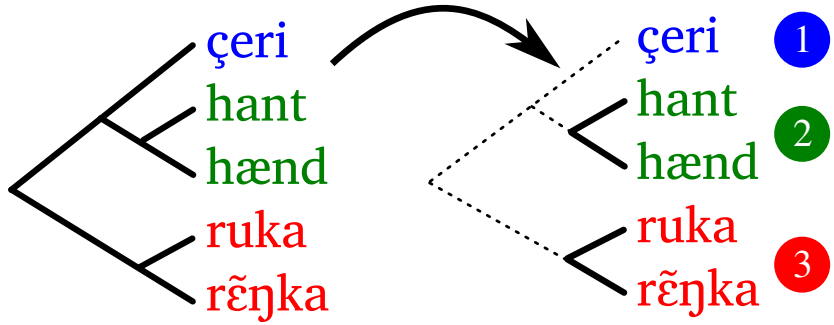
List et al. 2017

LexStat: Cognate Set Partitioning

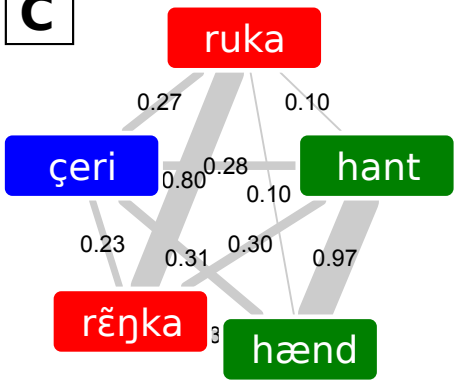
A

		GREEK	GERMAN	ENGLISH	RUSSIAN	POLISH
GREEK	çeri	0.00	0.72	0.69	0.73	0.77
GERMAN	hant	0.72	0.00	0.03	0.91	0.70
ENGLISH	hænd	0.69	0.03	0.00	0.91	0.68
RUSSIAN	ruka	0.72	0.91	0.91	0.00	0.20
POLISH	rɛ̃ŋka	0.77	0.70	0.68	0.20	0.00

B



C



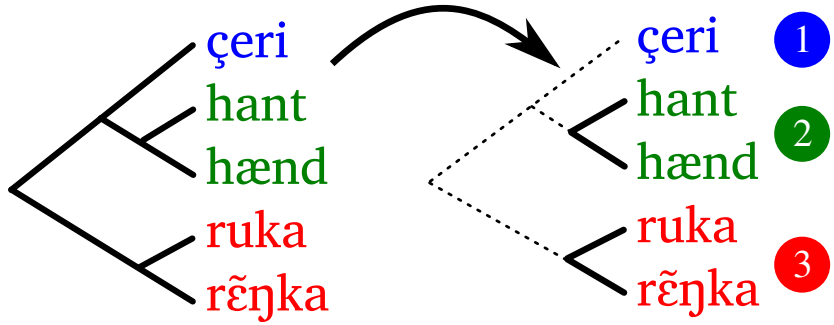
List et al. 2017

LexStat: Cognate Set Partitioning

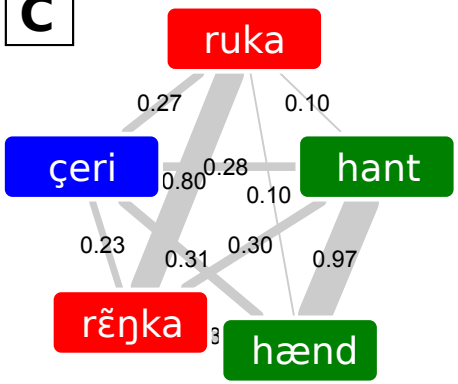
A

		GREEK	GERMAN	ENGLISH	RUSSIAN	POLISH
GREEK	çeri	0.00	0.72	0.69	0.73	0.77
GERMAN	hant	0.72	0.00	0.03	0.91	0.70
ENGLISH	hænd	0.69	0.03	0.00	0.91	0.68
RUSSIAN	ruka	0.72	0.91	0.91	0.00	0.20
POLISH	rɛ̃ŋka	0.77	0.70	0.68	0.20	0.00

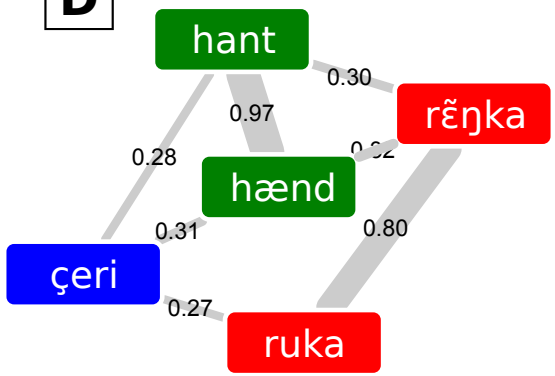
B



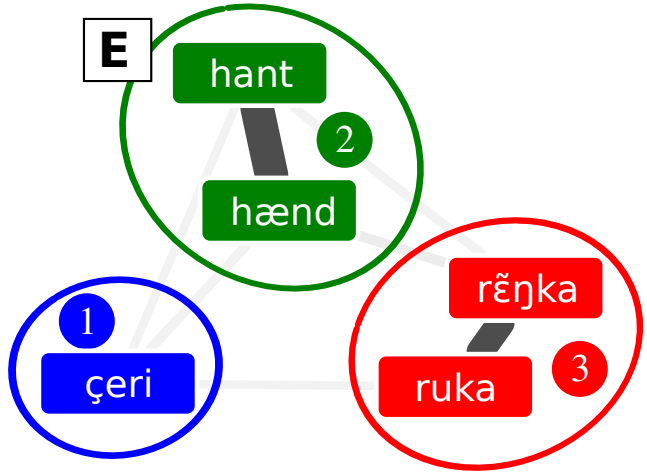
C



D



E



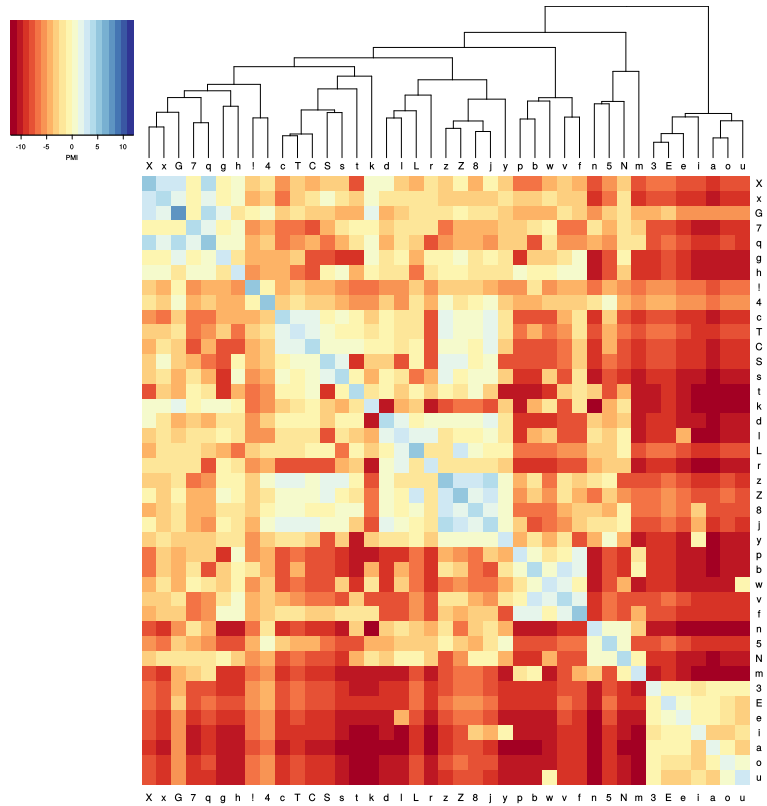
List et al. 2017

PMI string similarity

- *Pointwise Mutual Information* (PMI) between two sound classes a and b:

$$\mathrm{PMI}(a,b) \doteq \log \frac{P(a,b \text{ are homologous})}{P(a)P(b)}$$

- automatically trained from ASJP data (Jäger, 2013)
- PMI similarity between two strings: aggregate PMI score for optimal pairwise alignment of those strings



Calibrated PMI similarity

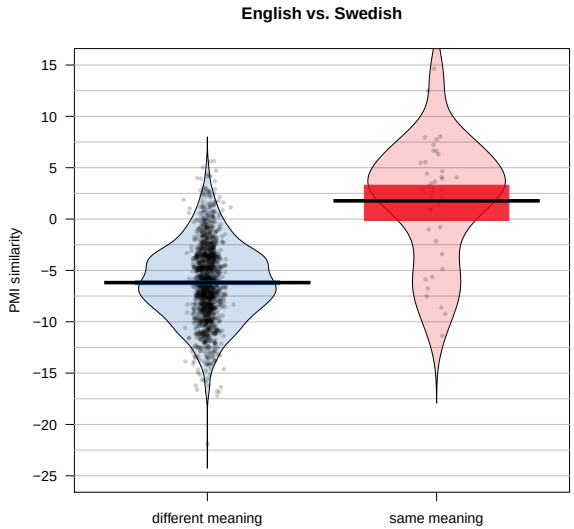
English / Swedish

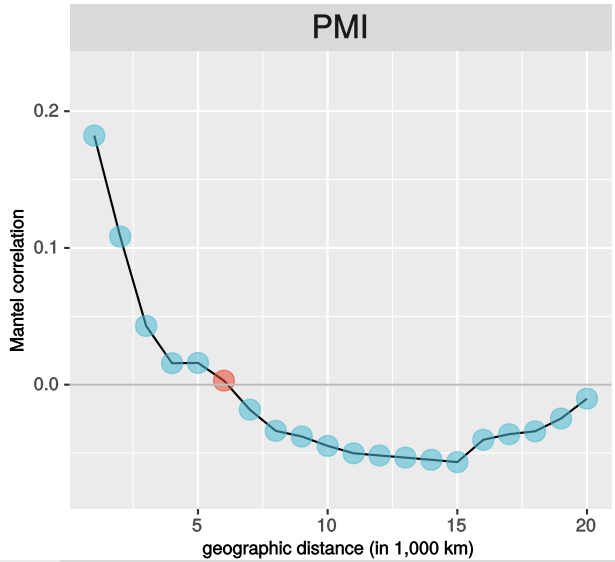
	Ei	yu	wi	w3n	tu	fiS	...
yog	**−7.77**	0.75	−7.68	−7.90	−8.57	−10.50	
du	−7.62	**0.33**	−5.71	−7.41	2.66	−8.57	
vi	−2.72	−2.83	**4.04**	−1.34	−6.45	0.70	
et	−5.47	−7.87	−5.47	**−6.43**	−1.83	−4.70	
tvo	−7.91	−4.27	−3.64	−4.57	**0.39**	−6.98	
fisk	−7.45	−11.2	−3.07	−9.97	−8.66	**7.58**	
⋮							

- values along diagonal give similarity between candidates for cognacy (possibility of meaning change is disregarded)
- values off diagonal provide sample of similarity distribution between non-cognates

Calibrated string similarity and language similarity

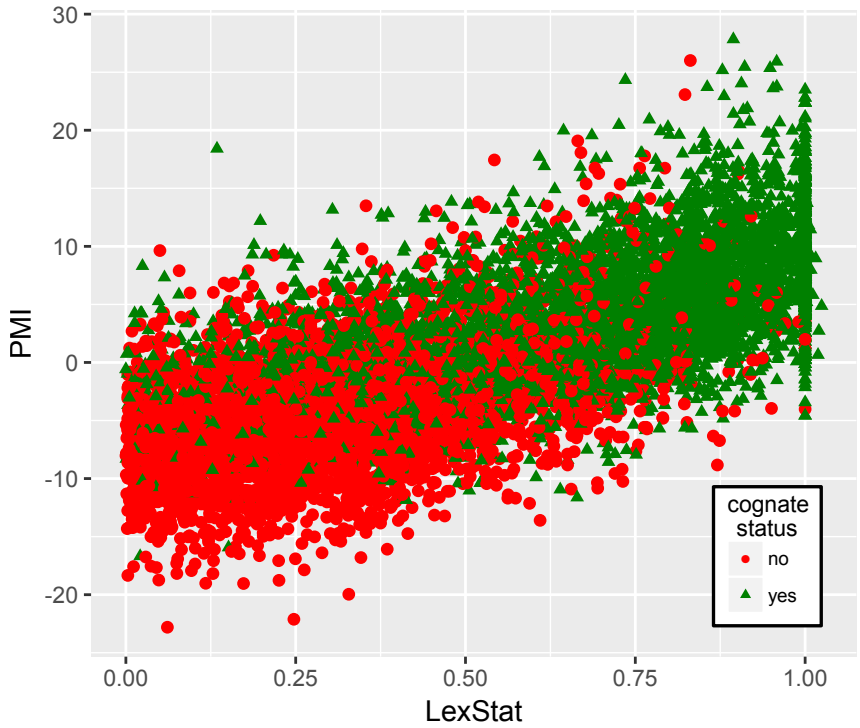
- let s be the PMI-similarity between the English and Swedish word for concept c
- **calibrated string similarity**: $-\log$(probability that random word pairs are more similar than s)
- **language similarity:** average word similarity for all concepts

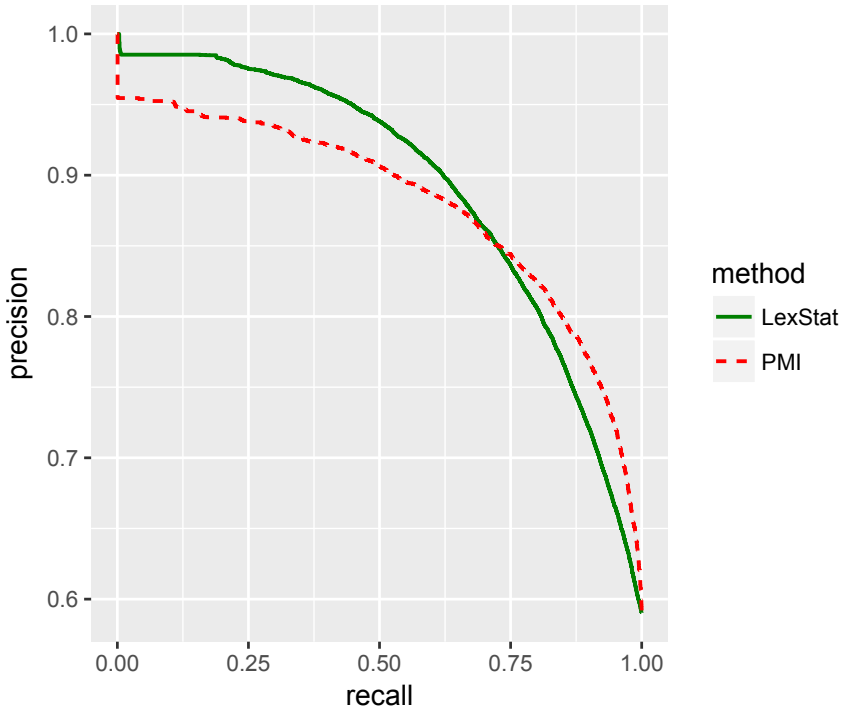




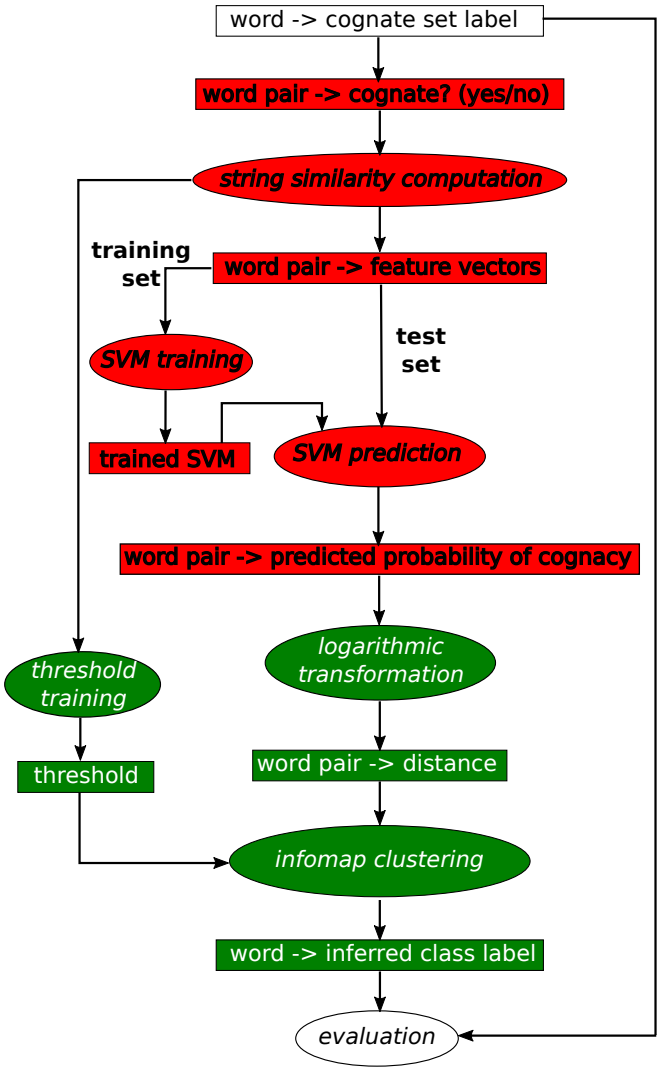
Comparing string similarity measures

- even though PMI similarity and LexStat similarity are based on different training methods, they capture a similar signal
- correlation: 0.727
- average precision LexStat: 0.893, PMI: 0.880

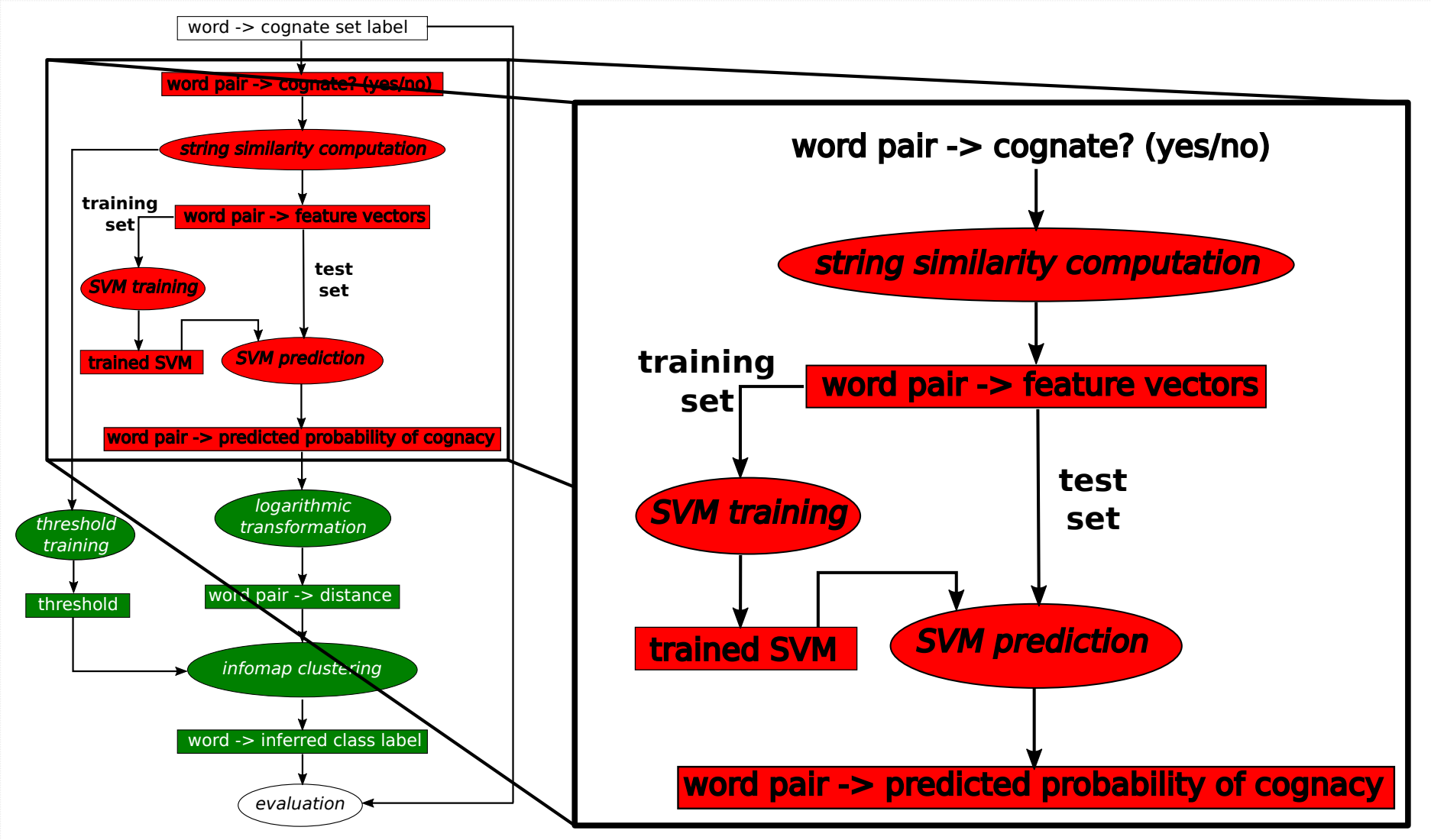




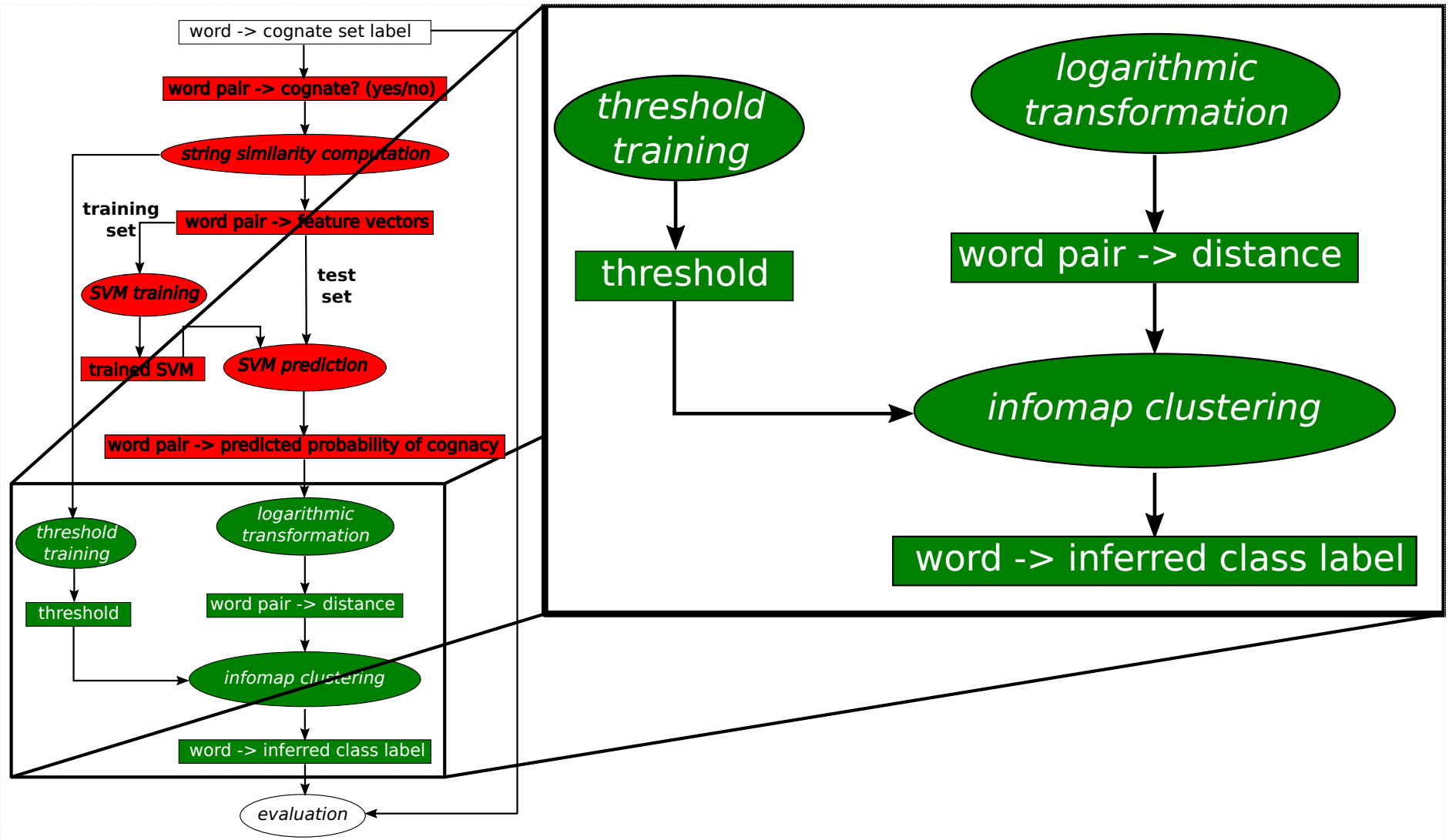
Workflow


word -> cognate set label
word pair -> cognate? (yes/no)
string similarity computation
training set
word pair -> feature vectors
test set
SVM training
trained SVM
SVM prediction
word pair -> predicted probability of cognacy
threshold training
logarithmic transformation
threshold
word pair -> distance
infomap clustering
word -> inferred class label
evaluation

Workflow


word -> cognate set label
word pair -> cognate? (yes/no)
string similarity computation
training set
word pair -> feature vectors
test set
SVM training
trained SVM
SVM prediction
word pair -> predicted probability of cognacy
threshold training
threshold
logarithmic transformation
word pair -> distance
infomap clustering
word -> inferred class label
evaluation
word pair -> cognate? (yes/no)
string similarity computation
training set
word pair -> feature vectors
test set
SVM training
trained SVM
SVM prediction
word pair -> predicted probability of cognacy

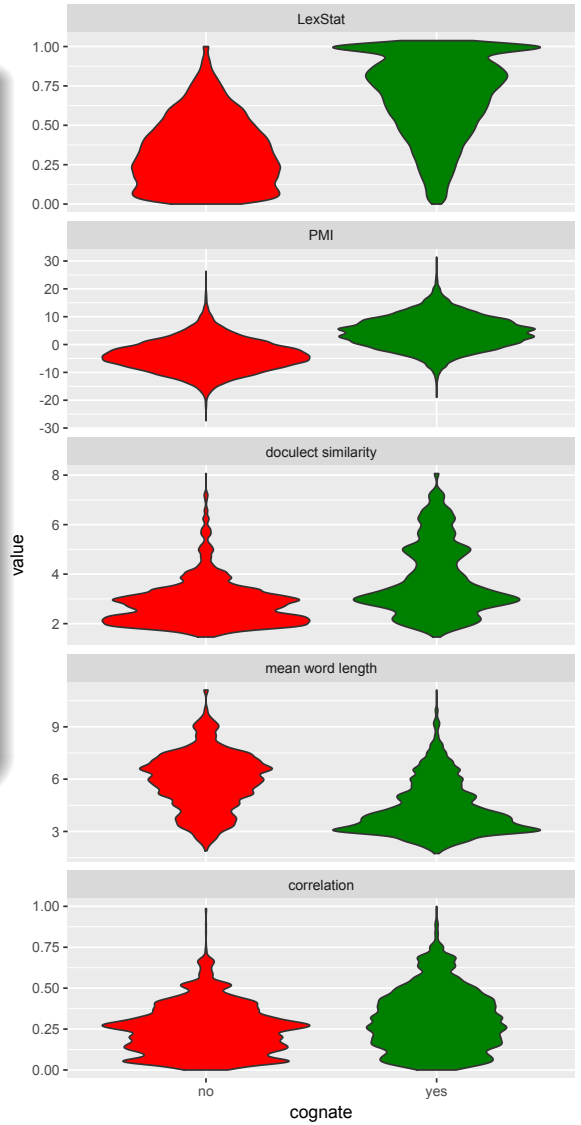
Workflow


word -> cognate set label
word pair -> cognate? (yes/no)
string similarity computation
training set
word pair -> feature vectors
test set
SVM training
trained SVM
SVM prediction
word pair -> predicted probability of cognacy
threshold training
threshold
logarithmic transformation
word pair -> distance
infomap clustering
word -> inferred class label
evaluation
threshold training
threshold
logarithmic transformation
word pair -> distance
infomap clustering
word -> inferred class label

SVM training

Model selection

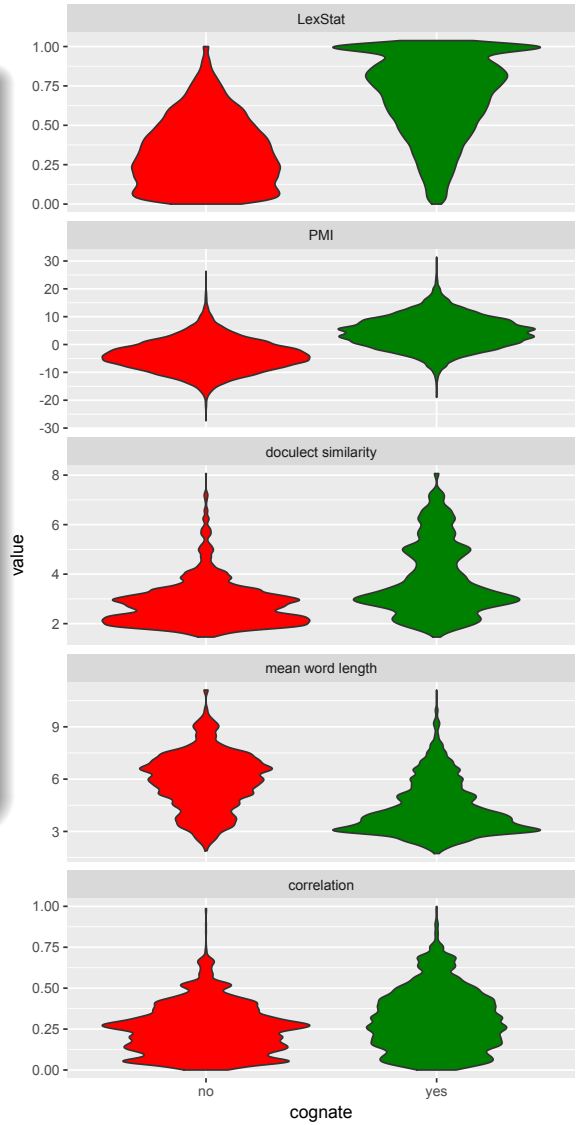
- each synonymous word pair is a data point
- cognate (yes/no) as dependent variable
- Feature selection
  - seven features from (Jäger and Sofroniev, 2016) + LexStat similarity as candidate features
  - feature selection via cross-validation on training data



SVM training

Model selection

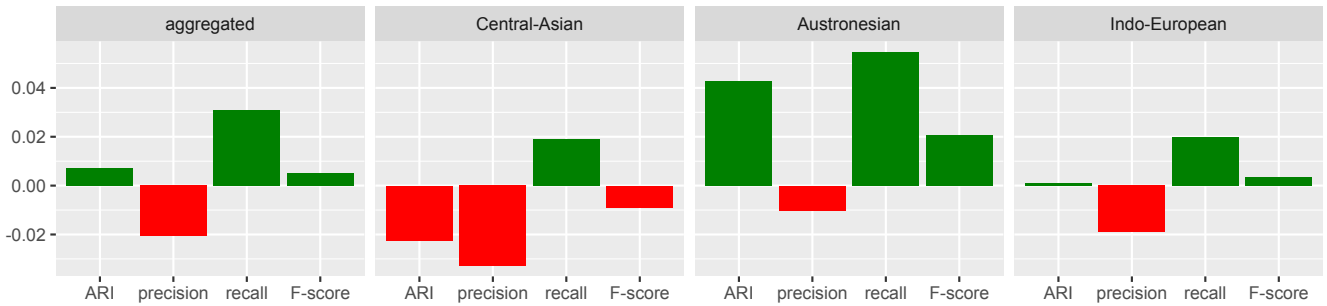
- five informative features
  - LexStat similarity
  - PMI similarity
  - doculect similarity
  - measures of concept stability
    - mean word length
    - correlation between string similarity and doculect similarity
- linear kernel



Evaluation

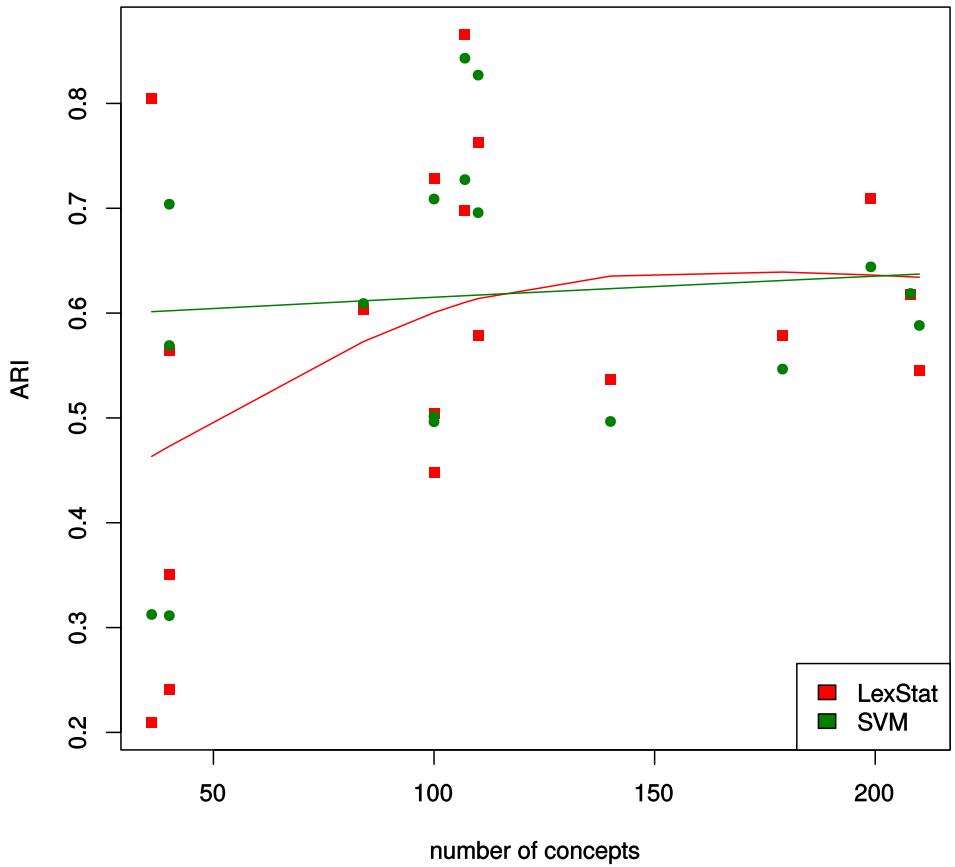
- two evaluation measures:
  - B-Cubed scores (Bagga and Baldwin, 1998)
  - Adjusted Rand Index (ARI, Hubert and Arabie 1985)
- LexStat clustering as benchmark

data set	*Adjusted Rand Index*		*B-Cubed Precision*		*B-Cubed Recall*		*B-Cubed F-Score*	
	LexStat	SVM	LexStat	SVM	LexStat	SVM	LexStat	SVM
aggregated	0.676	**0.683**	**0.868**	0.847	0.838	**0.869**	0.850	**0.855**
Austronesian	0.545	**0.588**	**0.791**	0.781	0.801	**0.855**	0.796	**0.817**
Central Asian	**0.866**	0.843	**0.916**	0.883	0.962	**0.981**	**0.938**	0.929
Indo-European	0.618	**0.619**	**0.896**	0.877	0.750	**0.770**	0.817	**0.820**



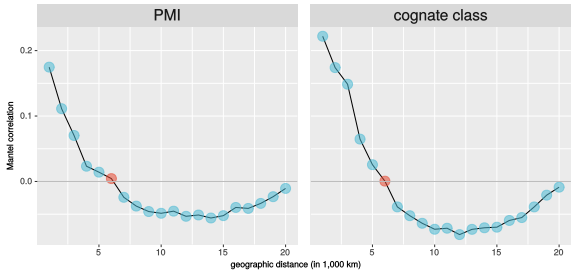
Evaluation

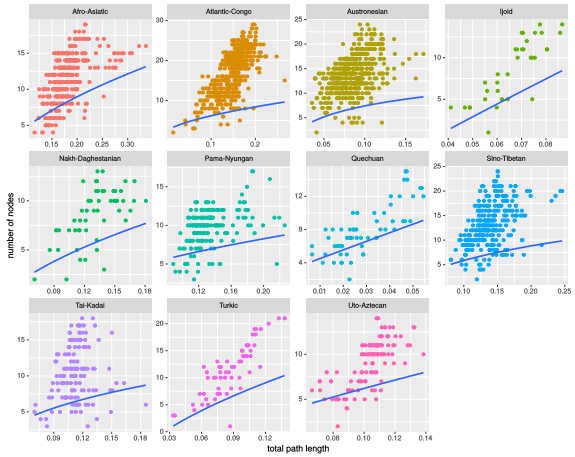
- overall, incremental improvement over LexStat
- however, massive improvement for short and low-quality word lists

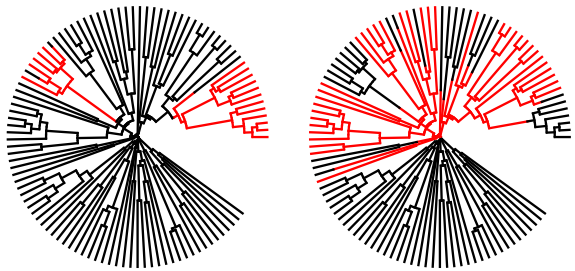


Outlook

- current methods for automatic cognate detection are sufficiently accurate to serve as a starting point for manual cognate annotation
- they are objective and replicable
- they make it much easier to obtain large datasets for cross-linguistic historical language comparison
- deep cognates, like English *cow* and Latin *bos*, remain a problem for automatic approaches which are based on surface similarity







Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the ACL*, pages 79–85, 1998.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960, Aug 2012.

Cecil H Brown, Eric W Holman, Søren Wichmann, and Viveka Velupillai. Automated classification of the world s languages. *Language Typology and Universals*, 61(4):285–308, 2008.

Běijīng Dàxué. *Hànyǔ fāngyán cíhuì* [Chinese dialect vocabularies]. Wénzì Gǎigé, 1964.

Michael Cysouw, Søren Wichmann, and David Kamholz. A critique of the separation base method for genealogical subgrouping. *Journal of Quantitative Linguistics*, 13(2-3):225–264, 2006.

Michael Dunn. Indo-European lexical cognacy database (IELex). URL: http://ielex.mpi.nl/, 2012.

Simon J. Greenhill, Robert Blust, and Russell D. Gray. The Austronesian Basic Vocabulary Database. *Evolutionary Bioinformatics*, 4:271–283, 2008.

Rebecca Grollemund, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences*, 112(43): 13296–13301, 2015.

Shirō Hattori. Japanese dialects. In Henry M. Hoenigswald and Robert H. Langacre, editors, *Diachronic, areal and typological linguistics*, pages 368–400. Mouton, The Hague and Paris, 1973.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. ISSN 1432-1343. doi: 10.1007/BF01908075. URL http://dx.doi.org/10.1007/BF01908075.

Hóu Jīngyī. Xiàndài Hànyǔ fāngyán yīnkù [Phonological database of Chinese dialects]. CD-ROM, 2004.

Gerhard Jäger. Phylogenetic inference from word lists using weighted alignment with empirical determined weights. *Language Dynamics and Change*, 3(2):245–291, 2013.

Gerhard Jäger and Pavel Sofroniev. Automatic cognate classification with a Support Vector Machine. In Stefanie Dipper, Friedrich Neubarth, and Heike Zinsmeister, editors, *Proceedings of the 13th Conference on Natural Language Processing*, volume 16 of *Bochumer Linguistische Arbeitsberichte*, pages 128–134. Ruhr Universität Bochum, 2016.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. *Concepticon*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2016. URL: http://concepticon.clld.org.

Kenneth A. McElhanon. Preliminary observations on Huon Peninsula languages. *Oceanic Linguistics*, 6(1):1–45, 1967. ISSN 00298115, 15279421. URL http://www.jstor.org/stable/3622923.

A IU Militarev. *Towards the chronology of Afrasian (Afroasiatic) and its daughter families*. McDonald Institute for Archaelogical Research, Cambridge, 2000.

Ilia Peiros. Comparative linguistics in Southeast Asia. *Pacific Linguistics*, 142, 1998.

Joy Sanders and Arden G Sanders. Dialect survey of the Kamasau language. *Pacific Linguistics. Series A. Occasional Papers*, 56:

137, 1980.

George S. Starostin. Annotated Swadesh wordlists for the Tujia group. In George S. Starostin, editor, *The Global Lexicostatistical Database*. RGGU, Moscow, 2013. URL: http://starling.rinet.ru.

Feng Wang. *Comparison of languages in contact. The distillation method and the case of Bai*. Institute of Linguistics Academia Sinica, Taipei, 2006.

Mikhail Zhivlov. Annotated Swadesh wordlists for the Ob-Ugrian group. In George S. Starostin, editor, *The Global Lexicostatistical Database*. RGGU, Moscow, 2011. URL: http://starling.rinet.ru.