Clustering Dialect Varieties Based on Historical Sound Correspondences

Verena Blaschke

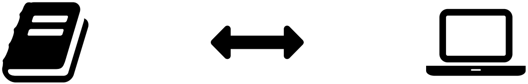
October 10th, 2019
GSCL Student Talks

Bachelor's Thesis
Supervised by Dr. Çağrı Çöltekin
Eberhard Karls Universität Tübingen

Introduction

Historical linguistics & dialectology

- Rarely been combined (lack of data)



- Efficiency / amount of data vs. time
- Feature selection

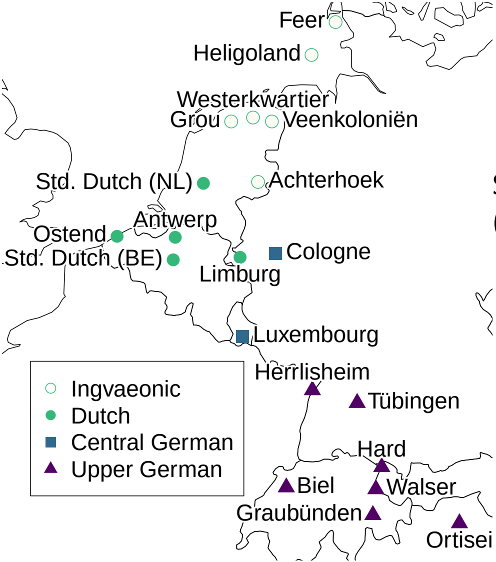
Introduction

My BA thesis:

- Modern continental West Germanic dialects
- Regular sound changes from Proto-Germanic

- Can we meaningfully cluster the dialects based on shared sound changes?
- Compare two clustering methods

Data

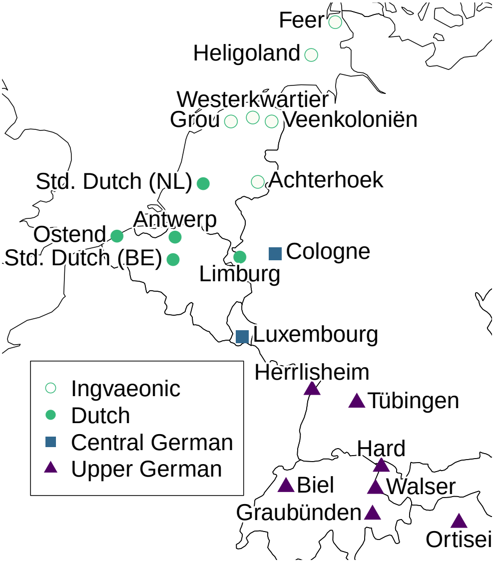


Sound Comparisons (Heggarty, 2018)

- 20 modern dialects
- Reconstruction of Proto-Germanic
- 110 cognate sets

Data

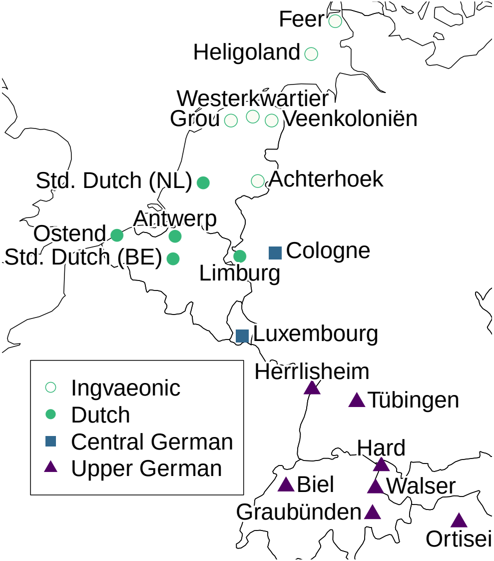
A Gold Standard?



- Ingvæonic (North Sea Germanic) vs. rest

Data

A Gold Standard?



- Ingvæonic (North Sea Germanic) vs. rest
- High German sound shift
  - /*p/ > /$\widehat{\text{pf}}$/ or /f/
  - /*t/ > /$\widehat{\text{ts}}$/ or /s/
  - /*k/ > /$\widehat{\text{kx}}$/ or /x/

(Upper, Central, Low DE + NL/Fris)

Method

1. Align segments 2. Deduce sound changes 3. Cluster dialects

Method

1. Align segments 2. Deduce sound changes 3. Cluster dialects

Dialect	Sound segments					
Proto-Germanic	k	a	l	d	a	z
Walser	x	aː	l	t	-	-
Biel	χ	ɑʊ	-	t	-	-
Luxembourg	kʰ	aː	l	-	-	-
...						

Progressive multiple sequence alignment
(Notredame et al., 2000)

Method

1. Align segments
2. Deduce sound changes
3. Cluster dialects

Dialect	Sound segments					
Proto-Germanic	k	a	l	d	a	z
Walser	x	aː	l	t	-	-
Biel	χ	ɑʊ	-	t	-	-
Luxembourg	k^h	aː	l	-	-	-
...						

Method

1. Align segments 2. Deduce sound changes 3. Cluster dialects

Dialect	Sound segments					
Proto-Germanic	k	a	l	d	a	z
Walser	x	aː	l	t	-	-
Biel	χ	ɑʊ	-	t	-	-
Luxembourg	k^h	aː	l	-	-	-
...						

k > x	no context
k > x / # _	word boundaries
k > x / _ vow	simple context (cons/vow)
k > x / _ A	sound class-based context (List, 2012)

Method

1. Align segments 2. Deduce sound changes 3. Cluster dialects

Dialect	Sound segments					
Proto-Germanic	k	a	l	d	a	z
Walser	x	aː	l	t	-	-
Biel	χ	ɑʊ	-	t	-	-
Luxembourg	k^h	aː	l	-	-	-
...						

$k > x$
$k > x \,/\, \#\,\_$
$k > x \,/\, \_\ \text{vow}$
$k > x \,/\, \_\ A$

$$
\begin{array}{c c}
 & \begin{array}{ccc} k > x & k > x \,/\, \#\_ & \cdots \end{array} \\
\begin{array}{c} \text{Walser} \\ \text{Biel} \\ \text{Lux.} \\ \text{Westk.} \\ \cdots \end{array} &
\begin{bmatrix} 1 & 1 & \cdots \\ 0 & 0 & \cdots \\ 0 & 0 & \cdots \\ 0 & 0 & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}
\end{array}
$$

Method

1. Align segments 2. Deduce sound changes 3. Cluster dialects

- Clustering only dialects vs. clustering dialects and sound correspondences at the same time
- Sound correspondences without/with phonetic context

Method

1. Align segments 2. Deduce sound changes 3. Cluster dialects

- Clustering only dialects vs. clustering dialects and sound correspondences at the same time
- Sound correspondences without/with phonetic context

Both clustering methods:

1. Normalize the dialect-by-correspondence tally matrix
2. Hierarchical clustering

Method

1. Align segments 2. Deduce sound changes 3. Cluster dialects

Clustering only dialects (agglomerative clustering)

1. Normalization: TF-IDF weighting
2. Clustering: Dialect-by-correspondence matrix ➔ dialect-by-dialect distance matrix ➔ dendrogram

Method

1. Align segments 2. Deduce sound changes 3. Cluster dialects

Clustering only dialects (agglomerative clustering)

1. Normalization: TF-IDF weighting
2. Clustering: Dialect-by-correspondence matrix ➔ dialect-by-dialect distance matrix ➔ dendrogram

Clustering dialects and sound correspondences (bipartite spectral graph co-clustering (Dhillon, 2001; Wieling and Nerbonne, 2010))

1. Normalization: Map all dialects and sound correspondences to vectors in a shared vector space
2. Clustering: K-means clustering (k=2)
3. Repeat

Method

1. Align segments 2. Deduce sound changes 3. Cluster dialects

- Now we have clusters of dialects / dialects and sound correspondences
- Rank the sound correspondences in each cluster:
  - Representativeness
  - Distinctiveness

Results & Discussion

No context vs. context

- With context: Higher representativeness & distinctiveness scores
- High-ranking sound correspondences often with consonant/vowel or word boundary information

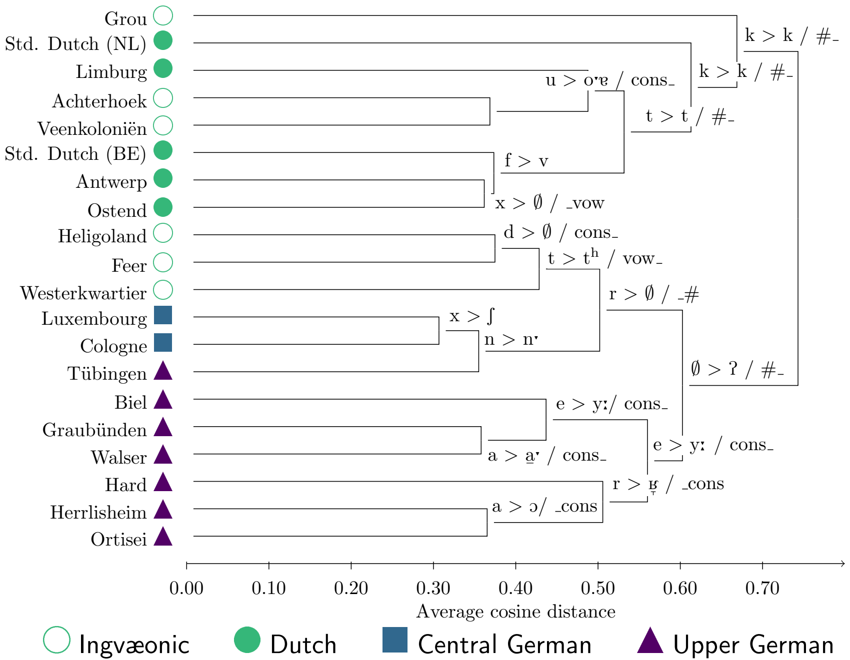
Results & Discussion

No context vs. context

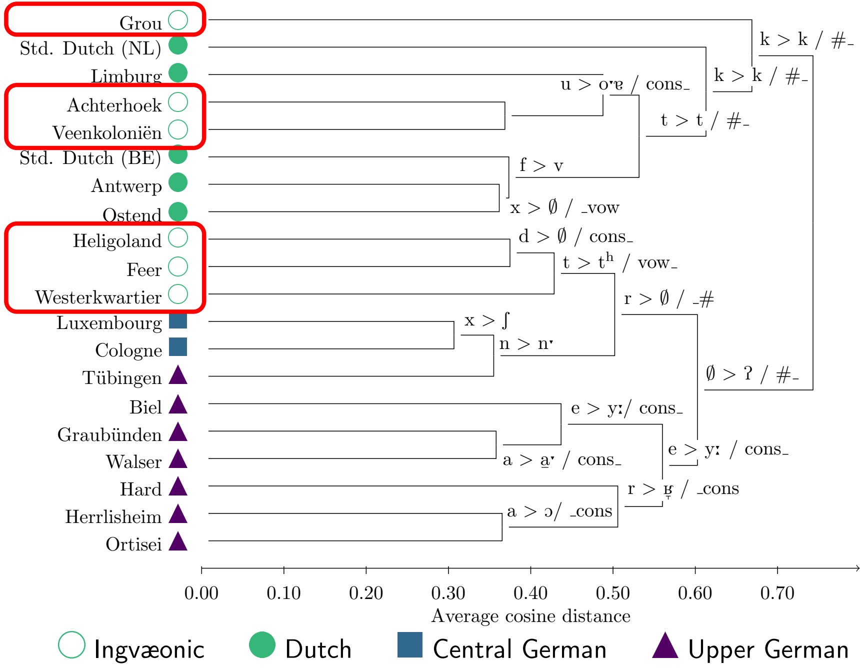
- With context: Higher representativeness & distinctiveness scores
- High-ranking sound correspondences often with consonant/vowel or word boundary information

Clustering methods

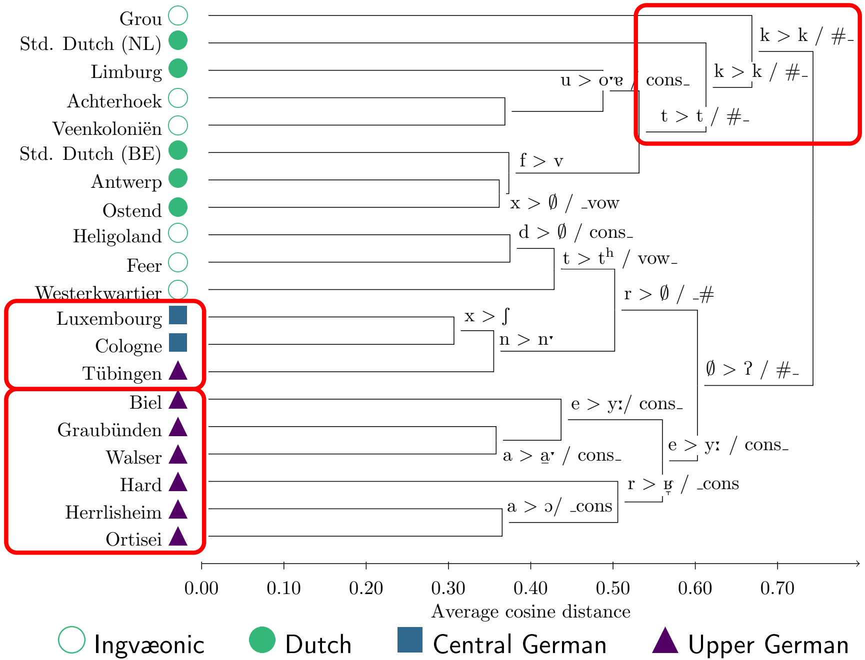
- Agglomerative clustering: More high-ranking sound correspondences, closer to the literature
- Graph co-clustering: Worked well in other researchers' dialectometry experiments – effects of data selection and pre-processing



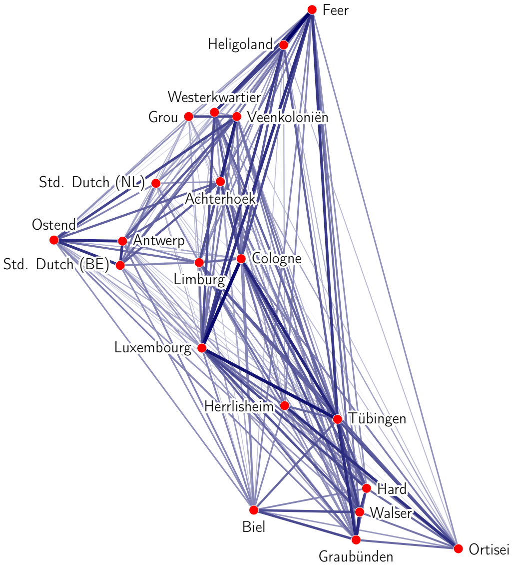
Grou
Std. Dutch (NL)
Limburg
Achterhoek
Veenkoloniën
Std. Dutch (BE)
Antwerp
Ostend
Heligoland
Feer
Westerkwartier
Luxembourg
Cologne
Tübingen
Biel
Graubünden
Walser
Hard
Herrlisheim
Ortisei
k > k / #_
k > k / #_
u > oˑɐ / cons_
t > t / #_
f > v
x > ∅ / _vow
d > ∅ / cons_
t > tʰ / vow_
r > ∅ / _#
x > ʃ
n > nˑ
∅ > ʔ / #_
e > yː/ cons_
e > yː / cons_
a > a̠ˑ / cons_
r > ʁ̞ / _cons
a > ɔ/ _cons
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
Average cosine distance
Ingvæonic
Dutch
Central German
Upper German



Grou
Std. Dutch (NL)
Limburg
Achterhoek
Veenkoloniën
Std. Dutch (BE)
Antwerp
Ostend
Heligoland
Feer
Westerkwartier
Luxembourg
Cologne
Tübingen
Biel
Graubünden
Walser
Hard
Herrlisheim
Ortisei
k > k / #_
k > k / #_
u > o'ɐ / cons_
t > t / #_
f > v
x > ∅ / _vow
d > ∅ / cons_
t > tʰ / vow_
r > ∅ / _#
x > ʃ
n > n'
∅ > ʔ / #_
e > yː/ cons_
e > yː / cons_
a > a' / cons_
r > ʁ / _cons
a > ɔ/ _cons
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
Average cosine distance
Ingvæonic
Dutch
Central German
Upper German

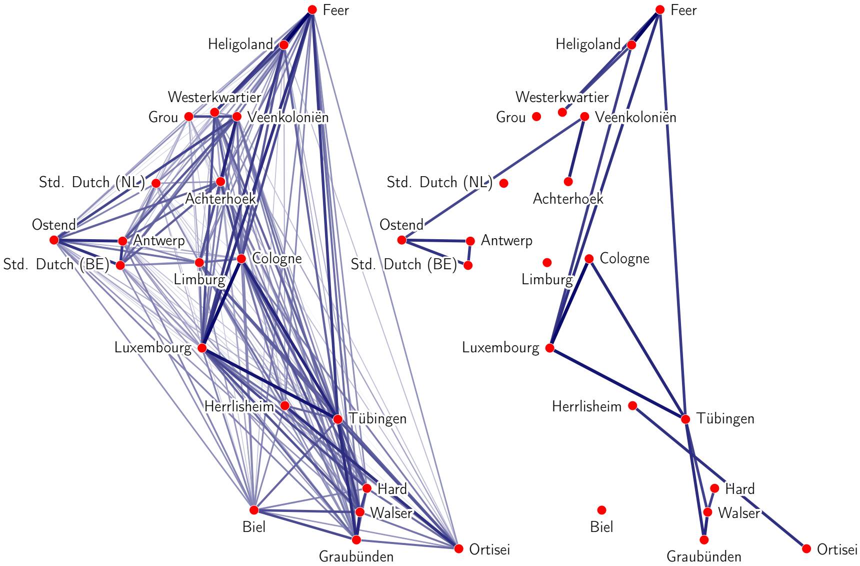

Grou
Std. Dutch (NL)
Limburg
Achterhoek
Veenkoloniën
Std. Dutch (BE)
Antwerp
Ostend
Heligoland
Feer
Westerkwartier
Luxembourg
Cologne
Tübingen
Biel
Graubünden
Walser
Hard
Herrlisheim
Ortisei
k > k / #_
k > k / #_
u > oˑɐ / cons_
t > t / #_
f > v
x > ∅ / _vow
d > ∅ / cons_
t > tʰ / vow_
r > ∅ / _#
x > ʃ
n > nˑ
∅ > ʔ / #_
e > yː/ cons_
a > a̠ˑ / cons_
e > yː / cons_
r > ʁ̞ / _cons
a > ɔ/ _cons
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
Average cosine distance
Ingvæonic
Dutch
Central German
Upper German

Results



Feer
Heligoland
Westerkwartier
Grou
Veenkoloniën
Std. Dutch (NL)
Achterhoek
Ostend
Antwerp
Std. Dutch (BE)
Limburg
Cologne
Luxembourg
Herrlisheim
Tübingen
Hard
Walser
Biel
Graubünden
Ortisei

Results


Feer
Heligoland
Westerkwartier
Grou
Veenkoloniën
Std. Dutch (NL)
Achterhoek
Ostend
Antwerp
Std. Dutch (BE)
Limburg
Cologne
Luxembourg
Herrlisheim
Tübingen
Hard
Walser
Biel
Graubünden
Ortisei
Feer
Heligoland
Westerkwartier
Grou
Veenkoloniën
Std. Dutch (NL)
Achterhoek
Ostend
Antwerp
Std. Dutch (BE)
Limburg
Cologne
Luxembourg
Herrlisheim
Tübingen
Hard
Walser
Biel
Graubünden
Ortisei

Conclusion

- We can infer meaningful structures
- Phonetic context is important!
- The simple approach (clustering only dialects) does better

`github.com/verenablaschke/dialect-clustering`

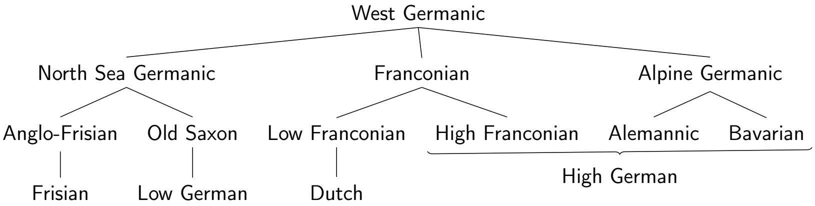
Bibliography I

- Dhillon, I. S. (2001). Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, San Francisco, pp. 269–274. ACM.
- Heggarty, P. (2018). Sound Comparisons: Exploring Diversity in Phonetics across Language Families. `www.soundcomparisons.com`
- List, J.-M. (2012). SCA: Phonetic Alignment Based on Sound Classes. In M. Slavkovik and D. Lassiter (Eds.), *New Directions in Logic, Language and Computation*, pp. 32–51. Berlin and Heidelberg: Springer.

Bibliography II

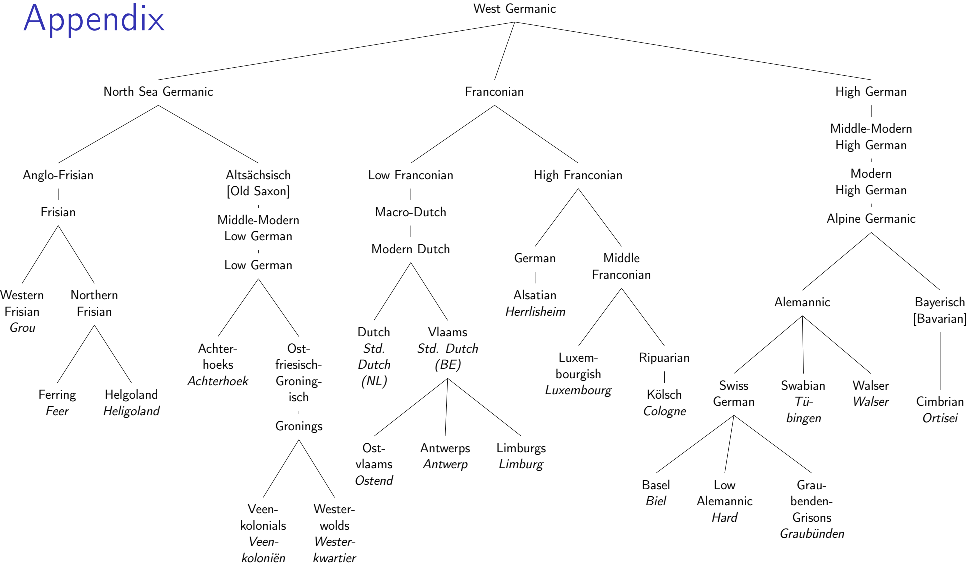
- Notredame, C., D. G. Higgins, and J. Heringa (2000). T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *Journal of Molecular Biology 302* (1), 205–217.
- Wieling, M. and J. Nerbonne (2011). Bipartite Spectral Graph Partitioning for Clustering Dialect Varieties and Detecting their Linguistic Features. *Computer Speech & Language 25* (3), 700– 715. Linguistics.

Appendix



Harbert, W. (2007). *The Germanic Languages*. Cambridge University Press, p. 8.

Appendix



Hammarström, H., R. Forkel, and M. Haspelmath (2018). Glottolog 3.3. Max Planck Institute for the Science of Human History.

Appendix

- 20 modern dialects
- 110 concepts in Proto-Germanic
- each modern dialect covers at least 103 concepts
- each concept is covered by at least 17 modern dialects
- 2181 word alignments between Proto-Germanic and modern dialects

Appendix

Ingvæonic / North Sea Germanic

- Frisian, English
- Maybe Low German?
- Maybe Dutch?

- Based on inflection and pronouns
- Sometimes based on phonological characteristics: palatalized velar consonants, fronted /a/, backed /a/ before nasals

Stiles, P. V. (2013). The Pan-West Germanic Isoglosses and the Subrelationships of West Germanic to Other Branches. *NOWELE. North-Western European Language Evolution 66* (1), 5–38.

Appendix

Progressive multiple sequence alignment with LingPy

1. Alignments for all possible pairwise combinations (modern or historical; Needleman-Wunsch algorithm)
2. Store aligned segments + frequencies in a library
3. Word-by-word distance matrix, convert into a tree (UPGMA)
4. Starting from the tips, progressively join 'sibling' alignments until all words are aligned at the root

Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994). CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research 22*, 4673–4680.

List, J.-M., S. Greenhill, and R. Forkel (2018). LingPy: A Python Library for Historical Linguistics. Version 2.6.3. With contributions by Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel, and Tiago Tresoldi.

Appendix

Abbr.	Definition	IPA characters
B	labial/labiodental fricatives	f, p͡f, v, ɸ, β
C	dental/alveolar affricates	d͡z, t͡s, t͡ʃ
D	dental fricatives	ð, θ
G	velar/uvual fricatives	x, χ, ɣ
H	laryngeals	h, ɦ, ʔ
J	palatal approximants	j
K	velar/uvular plosives/affricates	k, k͡x, q, g
L	lateral approximants	l, ɫ, ʟ
M	labial nasals	m, ɱ
N	(non-labial) nasals	n, ŋ, ɱ, ɴ
P	labial plosives	b, p
R	trills/taps/flaps	r, ɹ, ɾ, ʀ, ʁ
S	sibilant fricatives	s, z, ç, ʃ, ʒ, ʝ
T	dental/alveolar plosives	t, d, ʈ
W	labial approximants/fricatives	w

Appendix

Abbr.	Definition	IPA characters
A	unrounded open vowels	a, ɑ
E	unrounded mid vowels	e, æ, ɐ, ə, ɛ, ɜ, ʌ
I	unrounded close vowels	i, ɪ
O	rounded open vowels	ɒ
Y	rounded close vowels	u, y, ʊ, ʏ

List (2012)

Appendix

Pr.-G.	Ort.	No context	Simple context	Sound class-based context	Word boundaries
k	k^h	$k > k^h$	$k > k^h$ / _vow		$k > k^h$ / #_
a	ɔ	a > ɔ	a > ɔ/ cons_ a > ɔ / _con	a > ɔ/ K_ a > ɔ / _L	
l	l	l > l	l > l / vow_ l > l / _cons	l > l / A_	
d	t͡s	d > t͡s	d > t͡s / cons_	d > t͡s / L_	
a		a > ∅	a > ∅ / cons_		
z		z > ∅			z > ∅ / _#

Appendix

TF-IDF weighting

$$\text{tf}(\mathit{dialect}_i, \mathit{corres}_j) = \frac{\text{no. of occurrences of } \mathit{corres}_j \text{ in } \mathit{dialect}_i}{\text{no. of occurrences of all sound corres. in } \mathit{dialect}_i}$$

$$\text{idf}(\mathit{corres}_j) = \mathit{log}\left(\frac{\text{number of dialects}}{\text{number of dialects with } \mathit{corres}_j}\right)$$

$$\text{tf-idf}(\mathit{dialect}_i, \mathit{corres}_j) = \text{tf}(\mathit{dialect}_i, \mathit{corres}_j) \times (\text{idf}(\mathit{corres}_j) + 1)$$

Appendix

$$\text{cosine\_distance}(dialect_i, dialect_j) = 1 - \frac{dialect_i \cdot dialect_j}{\|dialect_i\| \|dialect_j\|}.$$

Unweighted Pair Group Method using Arithmetic Averages

1. Each dialect forms a singleton cluster.
2. Merge the two most similar clusters into a new cluster.
   Distance between this new cluster X and any given cluster Y:

$$\text{dist}(X, Y) = \sum_{x \in X} \sum_{y \in Y} \frac{\text{cosine\_distance}(x, y)}{|X| \times |Y|}.$$

3. Keep repeating step 2.

Sokal, R. R. and C. D. Michener (1958). A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin 28*, 1409–1438.

Appendix

Bipartite Spectral Graph Clustering (Dhillon, 2001; Wieling and Nerbonne, 2011)

1. Binary co-occurrence matrix $A \in \mathbb{R}^{m \times n}$
2. Normalization: Diagonal matrices $D_1 \in \mathbb{R}^{m \times m}$ and $D_2 \in \mathbb{R}^{n \times n}$ containing the row/column sums of A

$$A_n = D_1^{-\frac{1}{2}} \times A \times D_2^{-\frac{1}{2}}$$

3. SVD of A_n: $A_n = U \Sigma V^T$ ($U \in \mathbb{R}^{m \times m}$, diagonal matrix with values in descending order $\Sigma \in \mathbb{R}_{\geq 0}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$) to obtain the left and right singular vectors u_i and v_i
4. Pick the second singular vectors u_2, v_2 to calculate the vector $Z \in \mathbb{R}^{(m+n) \times 1}$:

$$Z_{[0,m]} = D_1^{-\frac{1}{2}} \times u_2; Z_{[m,m+n]} = D_2^{-\frac{1}{2}} \times v_2$$

5. K-means clustering on Z with $k = 2$.
6. Repeat for each cluster.

Appendix

Representativeness:

$$\text{rep}(X, A) = \frac{\text{number of dialects in } X \text{ with corres } A}{\text{number of dialects in } X}$$

Distinctiveness:

$$\text{relative\_occ}(X, A) = \frac{\text{no. of dialects in } X \text{ with corres} A}{\text{total no. of dialects with } A}$$

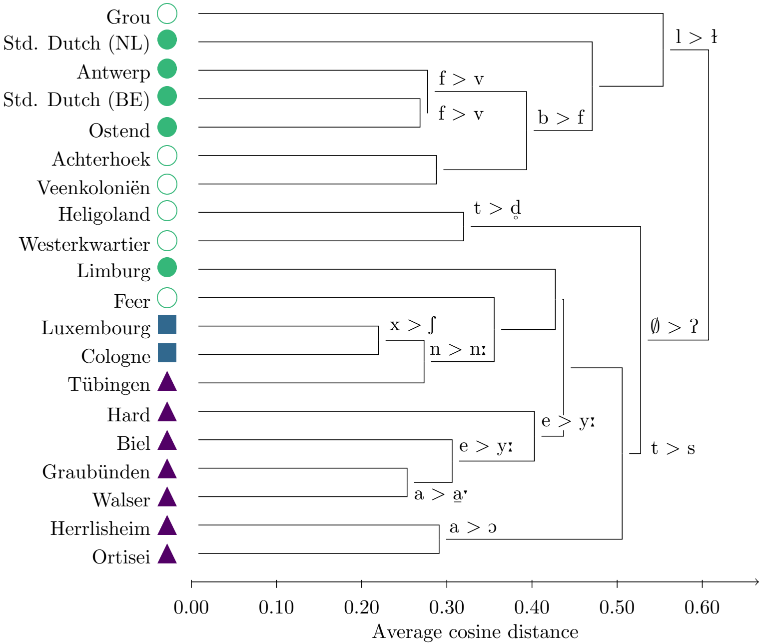
$$\text{relative\_size}(X) = \frac{\text{no. of dialects in } X}{\text{total no. of dialects}}$$

$$\text{dist}(X, A) = \frac{\text{relative\_occ}(X, A) - \text{relative\_size}()}{1 - \text{relative\_size}(X)}$$

Importance:

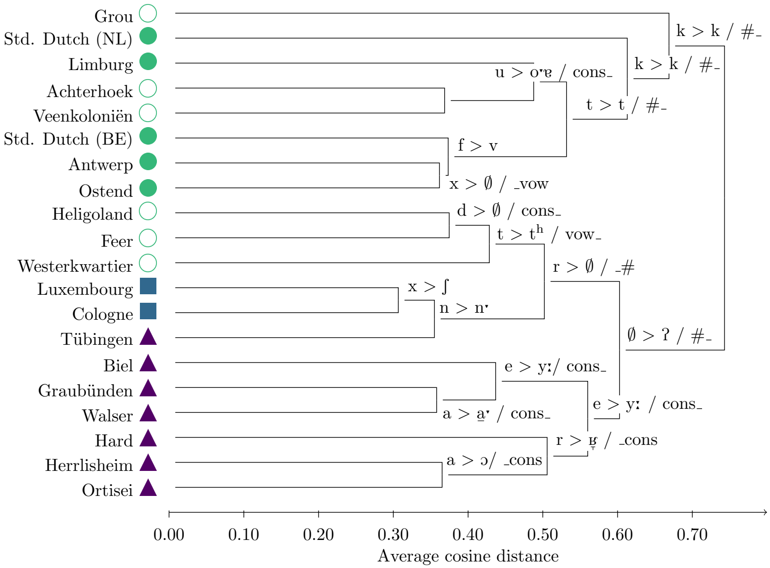
$$\text{imp}(X, A) = \begin{cases} \dfrac{2 * \text{rep}(X, A) * \text{dist}(X, A)}{\text{rep}(X, A) + \text{dist}(X, A)}, & \text{if dist}(X, A) > 0 \\ 0, & \text{otherwise} \end{cases}$$

Appendix



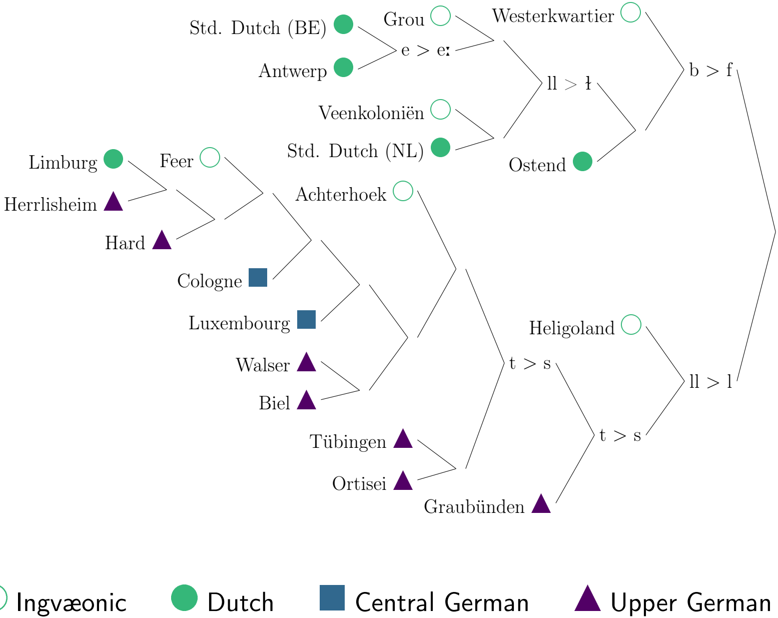
Grou
Std. Dutch (NL)
Antwerp
Std. Dutch (BE)
Ostend
Achterhoek
Veenkoloniën
Heligoland
Westerkwartier
Limburg
Feer
Luxembourg
Cologne
Tübingen
Hard
Biel
Graubünden
Walser
Herrlisheim
Ortisei
l > ɫ
f > v
f > v
b > f
t > d̥
∅ > ʔ
x > ʃ
n > nː
e > yː
e > yː
t > s
a > a̠ˑ
a > ɔ
0.00
0.10
0.20
0.30
0.40
0.50
0.60
Average cosine distance

Appendix



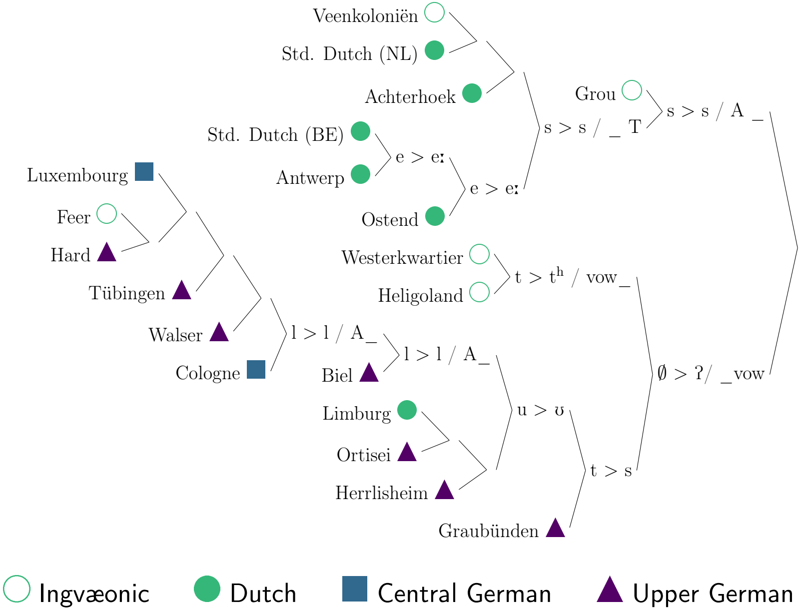
Grou
Std. Dutch (NL)
Limburg
Achterhoek
Veenkoloniën
Std. Dutch (BE)
Antwerp
Ostend
Heligoland
Feer
Westerkwartier
Luxembourg
Cologne
Tübingen
Biel
Graubünden
Walser
Hard
Herrlisheim
Ortisei
k > k / #_
k > k / #_
u > oˑɐ / cons_
t > t / #_
f > v
x > ∅ / _vow
d > ∅ / cons_
t > tʰ / vow_
r > ∅ / _#
x > ʃ
n > nˑ
∅ > ʔ / #_
e > yː/ cons_
e > yː / cons_
a > a̠ˑ / cons_
r > ʁ̞ / _cons
a > ɔ/ _cons
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
Average cosine distance

Appendix



Std. Dutch (BE)
Antwerp
Grou
e > eː
Westerkwartier
ll > ɫ
b > f
Veenkoloniën
Std. Dutch (NL)
Ostend
Limburg
Herrlisheim
Hard
Feer
Achterhoek
Cologne
Luxembourg
Walser
Biel
Tübingen
Ortisei
Graubünden
t > s
t > s
Heligoland
ll > l
Ingvæonic
Dutch
Central German
Upper German

Appendix


Veenkoloniën
Std. Dutch (NL)
Achterhoek
Grou
s > s / _ T
s > s / A _
Std. Dutch (BE)
Antwerp
e > eː
e > eː
Ostend
Luxembourg
Feer
Hard
Tübingen
Walser
Cologne
l > l / A_
Biel
l > l / A_
Westerkwartier
Heligoland
t > tʰ / vow_
Limburg
Ortisei
Herrlisheim
u > ʊ
Graubünden
t > s
∅ > ʔ/ _vow
Ingvæonic
Dutch
Central German
Upper German