Reconstructing Meaning Change from Parallel Corpora

Michael Cysouw and Jelena Prokić

Ludwig-Maximilians-Universität München

QITL-4, Berlin, March 31, 2011

Overview

1. Introduction
2. Experiment
3. Results
4. Further developments
5. Conclusions

Reconstructing history

- Sound changes ('historical-comparative method')

Reconstructing history

- Sound changes ('historical-comparative method')
- Changes in wordlists ('Swadesh approach')

Reconstructing history

- Sound changes ('historical-comparative method')
- Changes in wordlists ('Swadesh approach')
- Changes in grammatical structure ('typological approach')

Reconstructing history

- Sound changes ('historical-comparative method')
- Changes in wordlists ('Swadesh approach')
- Changes in grammatical structure ('typological approach')
- NEW: Changes in meaning

Definition of 'meaning'

- Define the meaning of a linguistic form as the set of all contexts in which it occurs

Definition of 'meaning'

- Define the meaning of a linguistic form as the set of all contexts in which it occurs
- Under this (strongly extensionalistic) definition of meaning, variation in meaning becomes readily measurable

Definition of 'meaning'

- Define the meaning of a linguistic form as the set of all contexts in which it occurs
- Under this (strongly extensionalistic) definition of meaning, variation in meaning becomes readily measurable
- Differences in meaning between cognates reveal historical processes

Generalizing wordlist comparison

- Wordlist approach:

Generalizing wordlist comparison

- Wordlist approach:
  - Select a set of 'meanings'

Generalizing wordlist comparison

- Wordlist approach:
  - Select a set of 'meanings'
  - Collect cognate forms expressing these meanings across different languages

Generalizing wordlist comparison

- Wordlist approach:
  - Select a set of 'meanings'
  - Collect cognate forms expressing these meanings across different languages
  - Non-cognates for the same meaning can be interpreted as the result of historical events

Generalizing wordlist comparison

- Wordlist approach:
  - Select a set of 'meanings'
  - Collect cognate forms expressing these meanings across different languages
  - Non-cognates for the same meaning can be interpreted as the result of historical events
- Generalization:

Generalizing wordlist comparison

- Wordlist approach:
  - Select a set of 'meanings'
  - Collect cognate forms expressing these meanings across different languages
  - Non-cognates for the same meaning can be interpreted as the result of historical events
- Generalization:
  - Select a large set of very similar contextual situations ('meanings')

Generalizing wordlist comparison

- Wordlist approach:
  - Select a set of 'meanings'
  - Collect cognate forms expressing these meanings across different languages
  - Non-cognates for the same meaning can be interpreted as the result of historical events
- Generalization:
  - Select a large set of very similar contextual situations ('meanings')
  - Collect the distribution of cognate forms over these contexts

Generalizing wordlist comparison

- Wordlist approach:
  - Select a set of 'meanings'
  - Collect cognate forms expressing these meanings across different languages
  - Non-cognates for the same meaning can be interpreted as the result of historical events
- Generalization:
  - Select a large set of very similar contextual situations ('meanings')
  - Collect the distribution of cognate forms over these contexts
  - Different distributions across languages can be interpreted as the result of historical events

Corpus

- Universal Declaration of Human Rights
- 13 Germanic languages
  - 6 prepositions: *in*, *under*, *with*, *through*, *for*, *against*
- 12 Slavic languages
  - 7 prepositions: *about*, *according to*, *before*, *for*, *from*, *in*, *on*

Distribution of the prepositions

- Find the distribution of each preposition within every paragraph in the text
- Example:

  These rights and freedoms may in no case be exercised contrary to the purposes and principles of the United Nations.

  Diese Rechte und Freiheiten dürfen in keinem Fall im Widerspruch zu den Zielen und Grundsätzen der Vereinten Nationen ausgeübt werden.

  Deze rechten en vrijheden mogen in geen geval worden uitgeoefend in strijd met de doeleinden en beginselen van de Verenigde Naties.

Distributional frequencies

- Relationships between the languages are inferred based on the distributional frequencies of the prepositions
- Example: Distribution of the preposition *in*

	Paragraph 1	Paragraph 2	Paragraph 3	Paragraph 4
German	1	1	0	0
English	1	2	0	0
Dutch	1	0	1	0
Frisian	2	1	2	1

Analyses of the frequencies

- Method 1

Analyses of the frequencies

- Method 1
  - distance between two languages is a sum of the absolute differences of the frequencies for each paragraph

Analyses of the frequencies

- Method 1
  - distance between two languages is a sum of the absolute differences of the frequencies for each paragraph
  - all pairwise distances are put into a single distance matrix

Analyses of the frequencies

- Method 1
  - distance between two languages is a sum of the absolute differences of the frequencies for each paragraph
  - all pairwise distances are put into a single distance matrix
  - the distances are analyzed using the neighbor-net algorithm (NN)

Analyses of the frequencies

- Method 1
  - distance between two languages is a sum of the absolute differences of the frequencies for each paragraph
  - all pairwise distances are put into a single distance matrix
  - the distances are analyzed using the neighbor-net algorithm (NN)
  - NN: produces a tree if the data is tree-like, and network is the data is network-like

Analyses of the frequencies

- Method 2

Analyses of the frequencies

- Method 2
  - analyze frequencies for each paragraph separately using the parsimony method

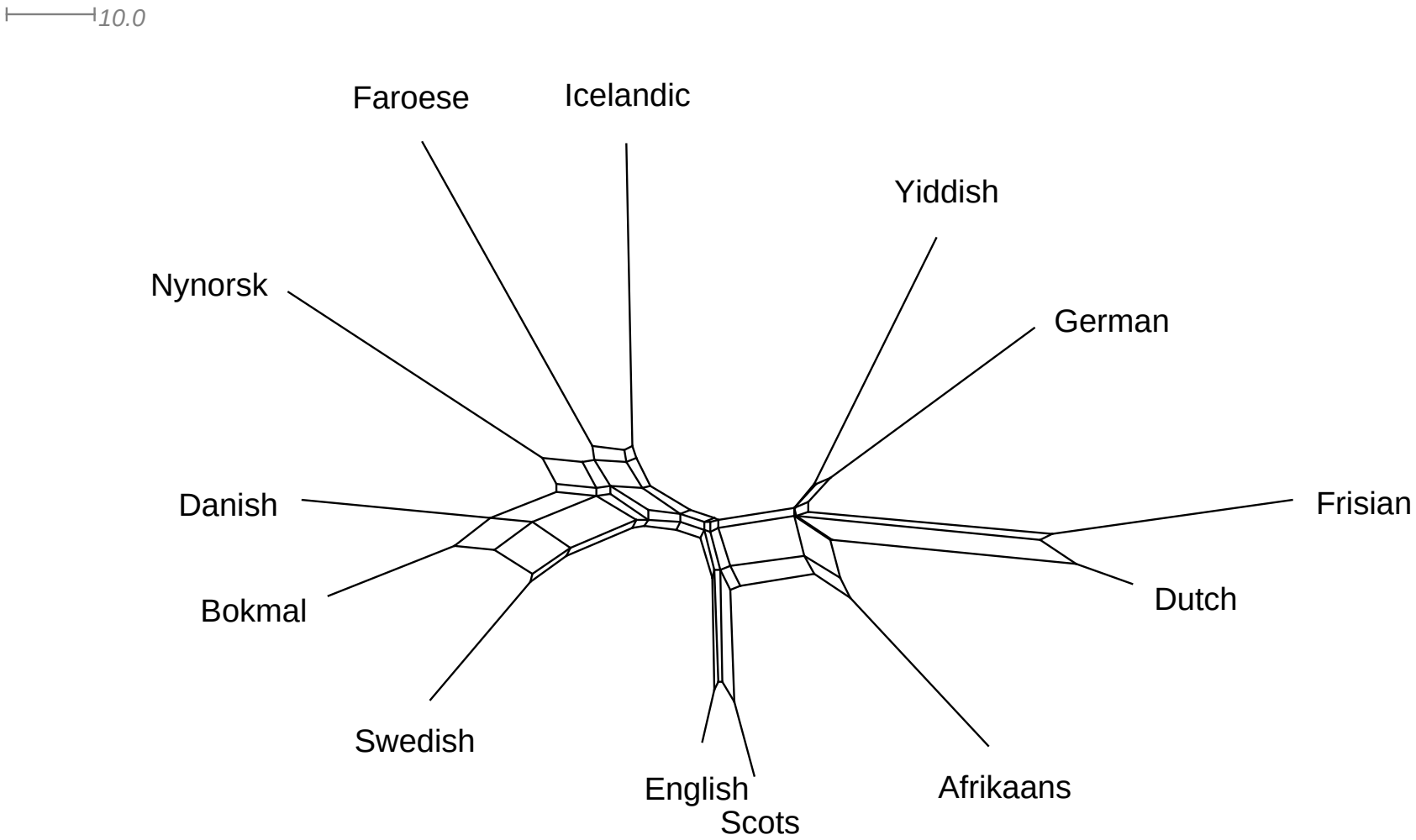
Analyses of the frequencies

- Method 2
  - analyze frequencies for each paragraph separately using the parsimony method
  - join all the analyses into a single tree

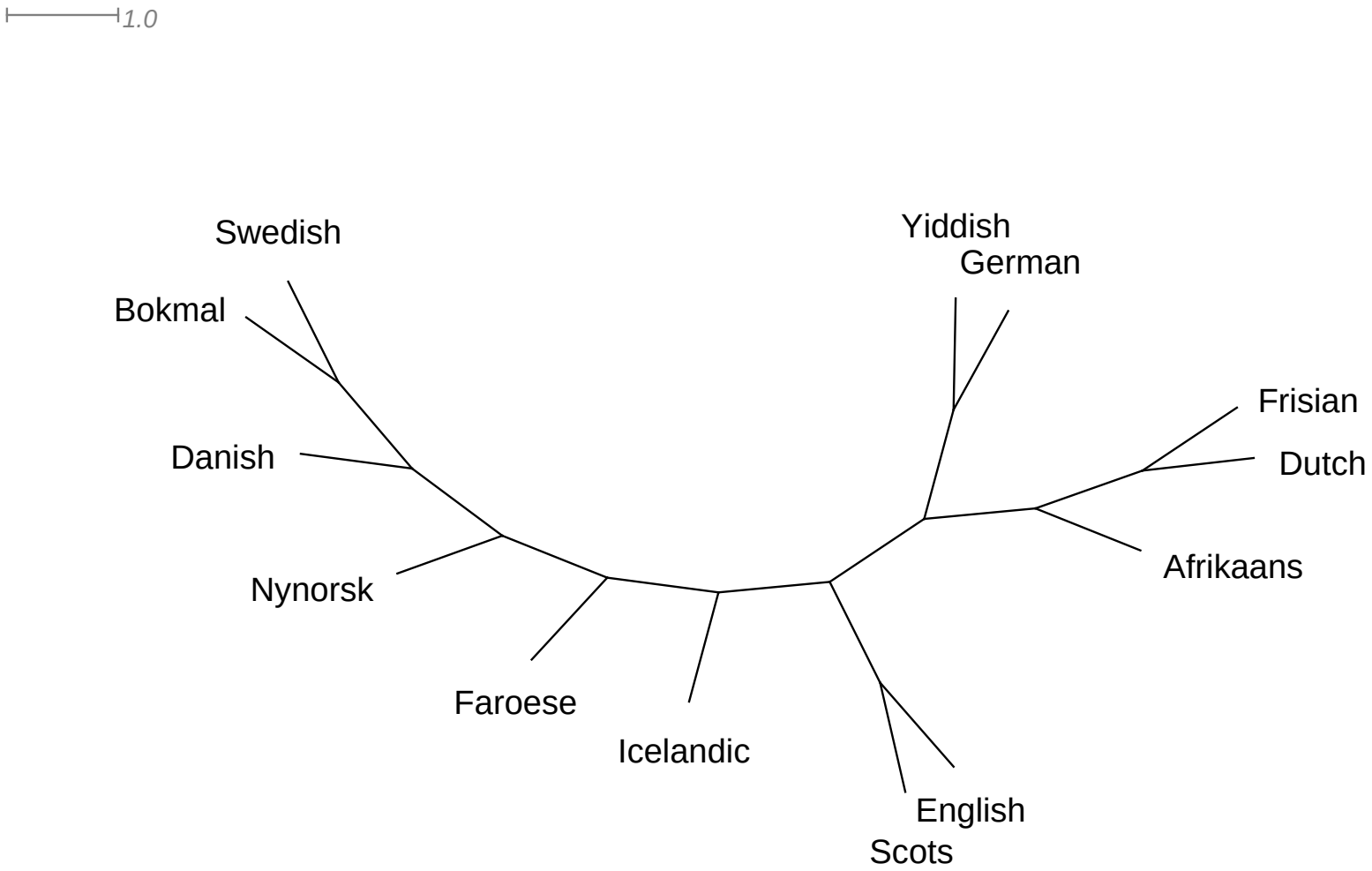
Analyses of the frequencies

- Method 2
  - analyze frequencies for each paragraph separately using the parsimony method
  - join all the analyses into a single tree
  - parsimony methods: search for the tree (grouping of the languages) that require the least amount of evolutionary change

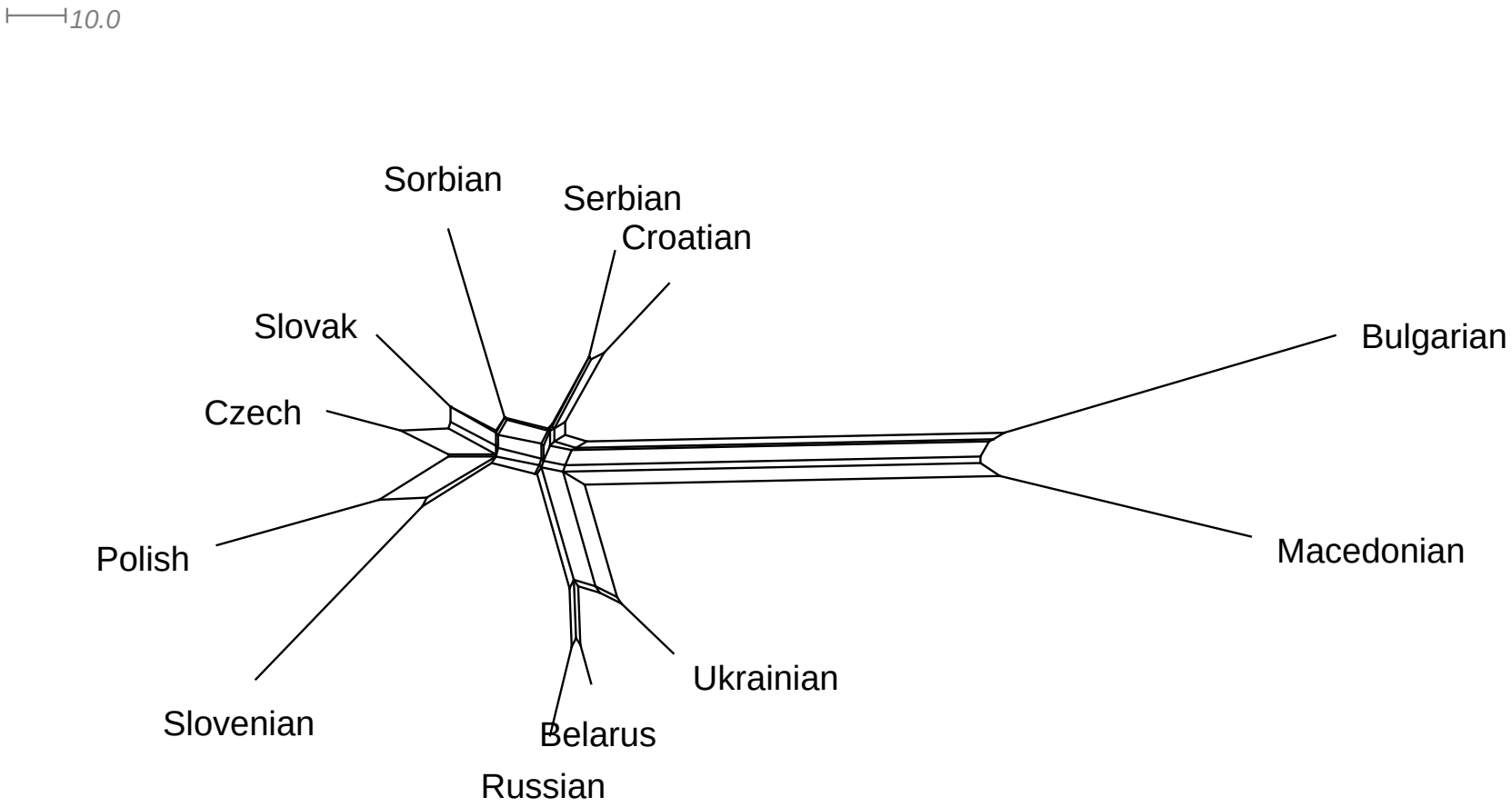
Germanic data: Neighbor-net



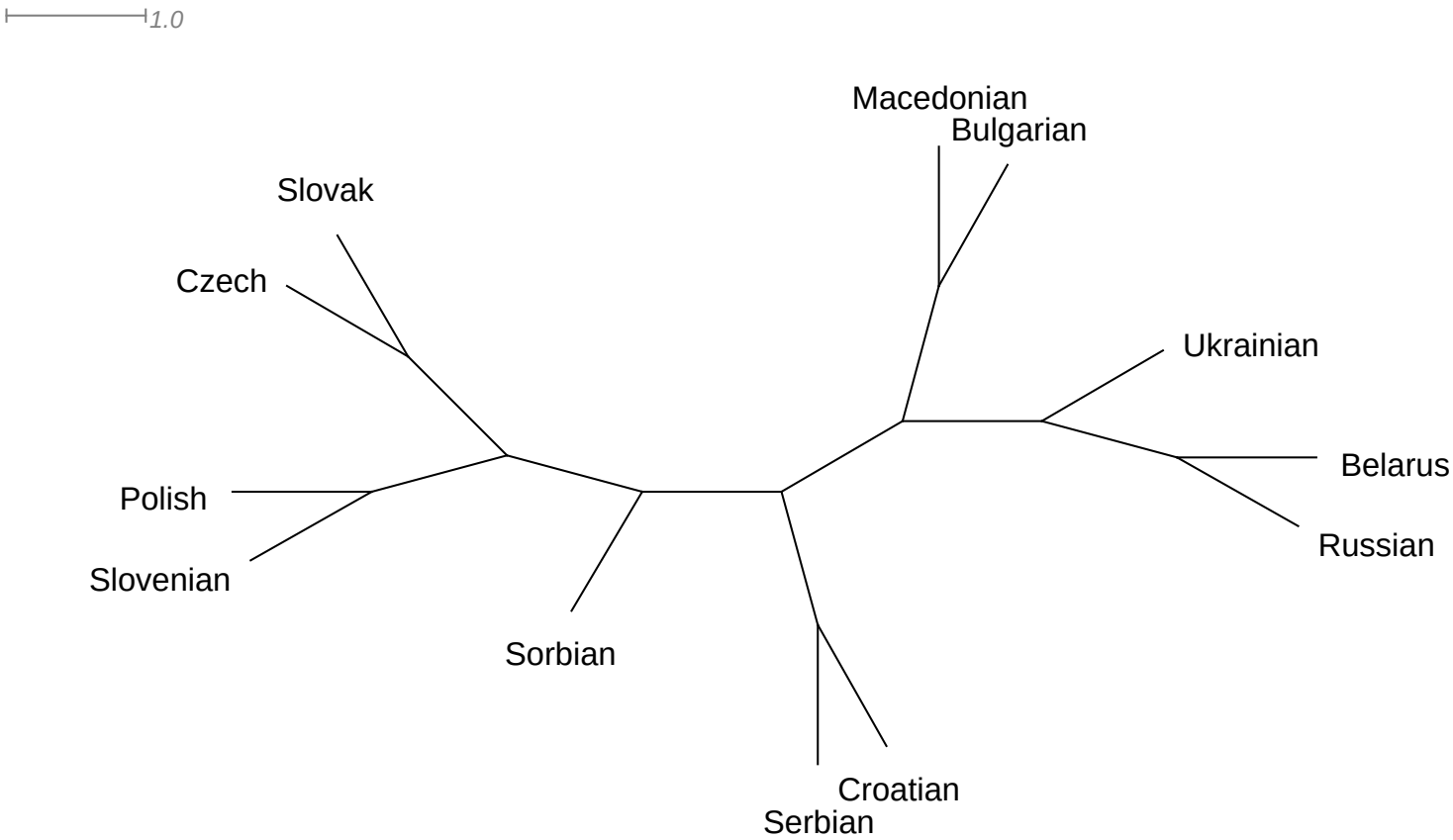
Germanic data: Maximum parsimony



Slavic data: Neighbor-net



Slavic data: Maximum parsimony



Proposal 1: Improve alignment

- Until now: count the number of prepositions per paragraph
- Better: try to align equivalent prepositions in the text

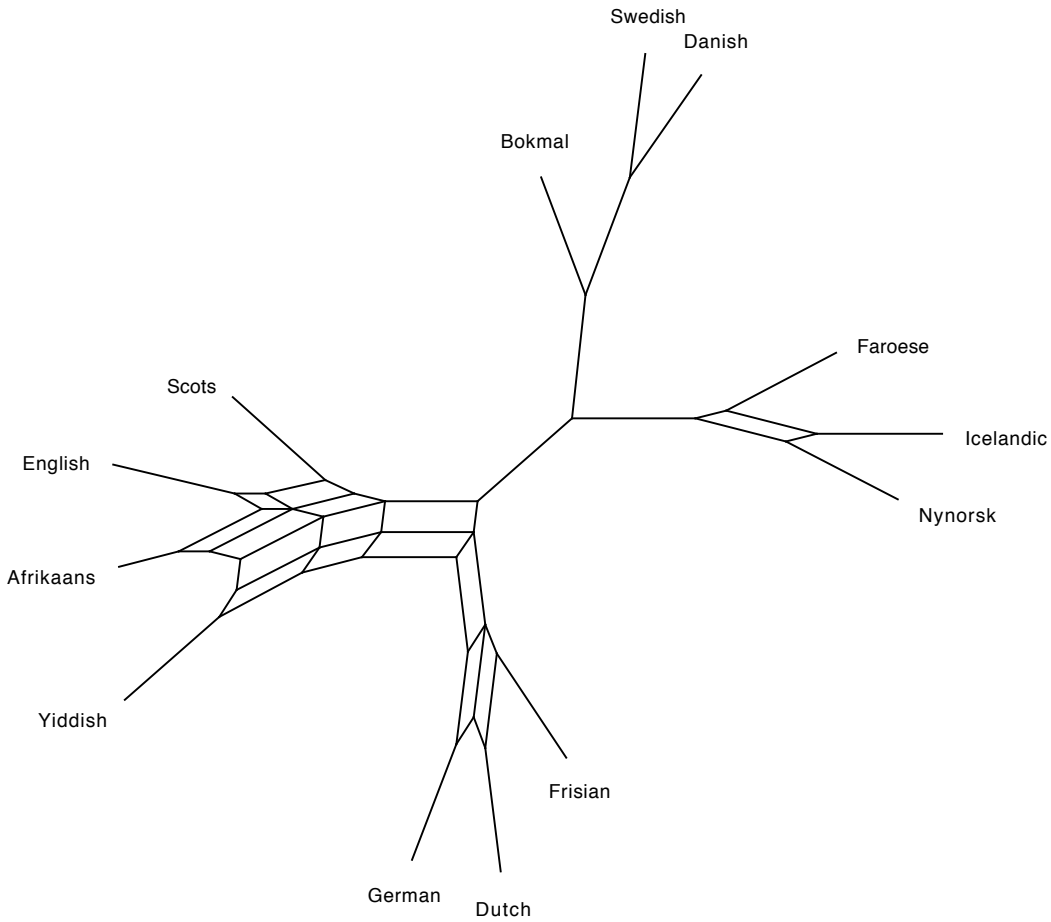
Preamble, paragraph 5

- English:
  Whereas the peoples of the United Nations have **in** the Charter reaffirmed their faith **in** fundamental human rights, **in** the dignity and worth of the human person and **in** the equal rights of men and women and have determined to promote social progress and better standards of life **in** larger freedom,
- German:
  da die Völker der Vereinten Nationen **in** der Charta ihren Glauben an die grundlegenden Menschenrechte, an die Würde und den Wert der menschlichen Person und an die Gleichberechtigung von Mann und Frau erneut bekräftigt und beschlossen haben, den sozialen Fortschritt und bessere Lebensbedingungen **in** grösserer Freiheit zu fördern,

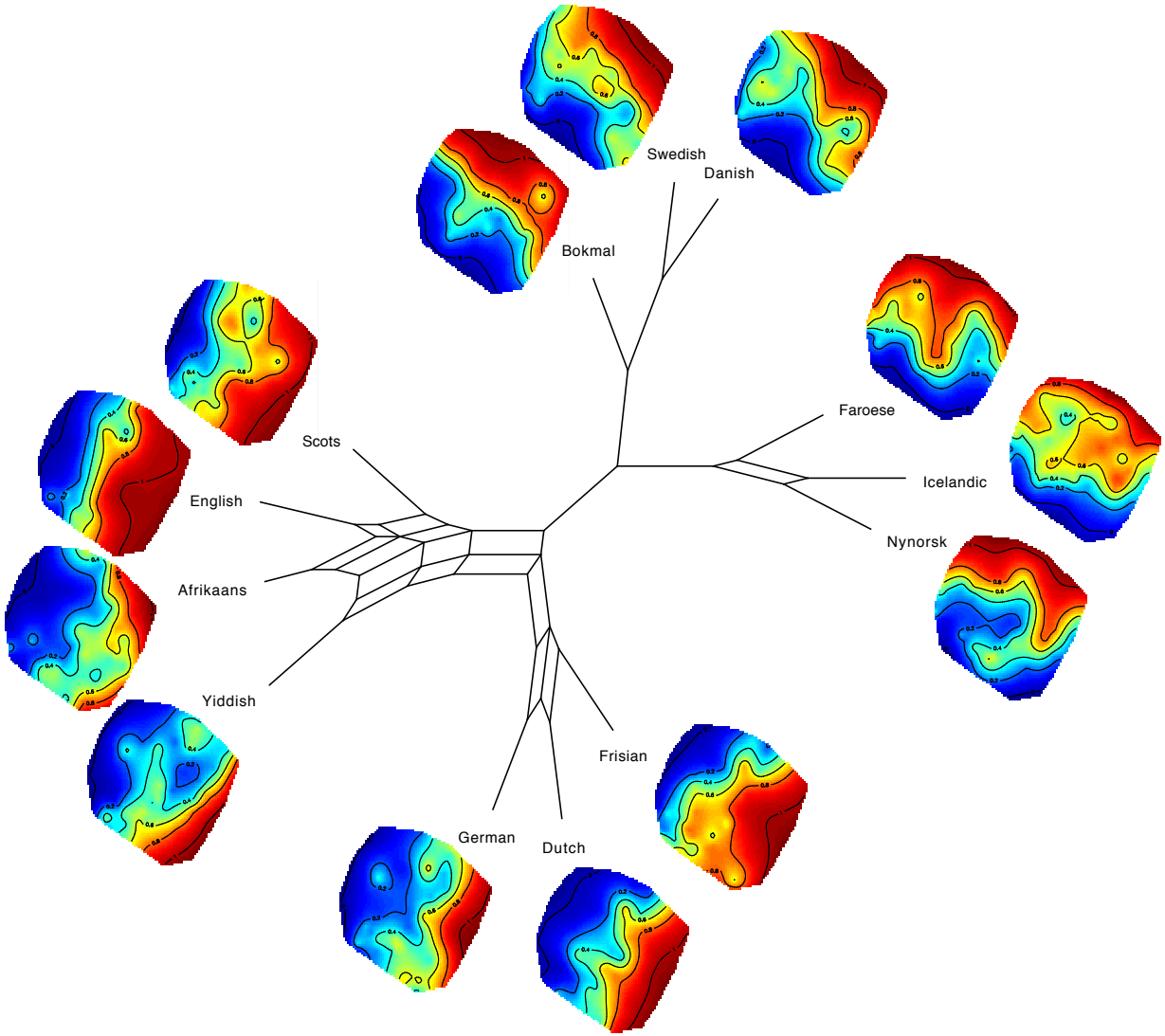
Language Frequency Positions in paragraph 5

Language	Frequency	Aligned					
German	2		1				1
English	5		1	1	1	1	1
Scots	5		1	1	1	1	1
Dutch	5		1	1	1	1	1
Frisian	6	1	1	1	1	1	1
Afrikaans	5		1	1	1	1	1
Yiddish	4		1	1	1		1
Nynorsk	3		1			1	1
Faroese	3		1			1	1
BokmI	1		1				
Swedish	1		1				
Danish	1		1				
Icelandic	1						1

Consensus network of 4 optimal trees according to dollo Maximum Parsimony



Differences in the distribution of 'in'



Proposal 2: Generalize this approach even further

- Define meaning of 'in' by the collection of words in its context
- Compare contextual vectors across languages

Step 1: Language specific contextual vectors

English	Frequency	German	Frequency
the	15	der	7
of	8	verkündeten	3
and	5	mit	3
with	4	Erklärung	3
which	4	dieser	3
this	4	die	3
rights	4	den	3

Step 2: Link words across languages

- Use the parallelism between the languages to estimate translational probability

$$sig(w, v) = \lambda^k \cdot e^{-\lambda} \cdot k!$$

$$\lambda = \frac{\text{frequency of word w} \cdot \text{frequency of word v}}{\text{number of contexts}}$$

k = frequency of co-occurrence of words w and v

Step 3: Compare contextual vectors across languages

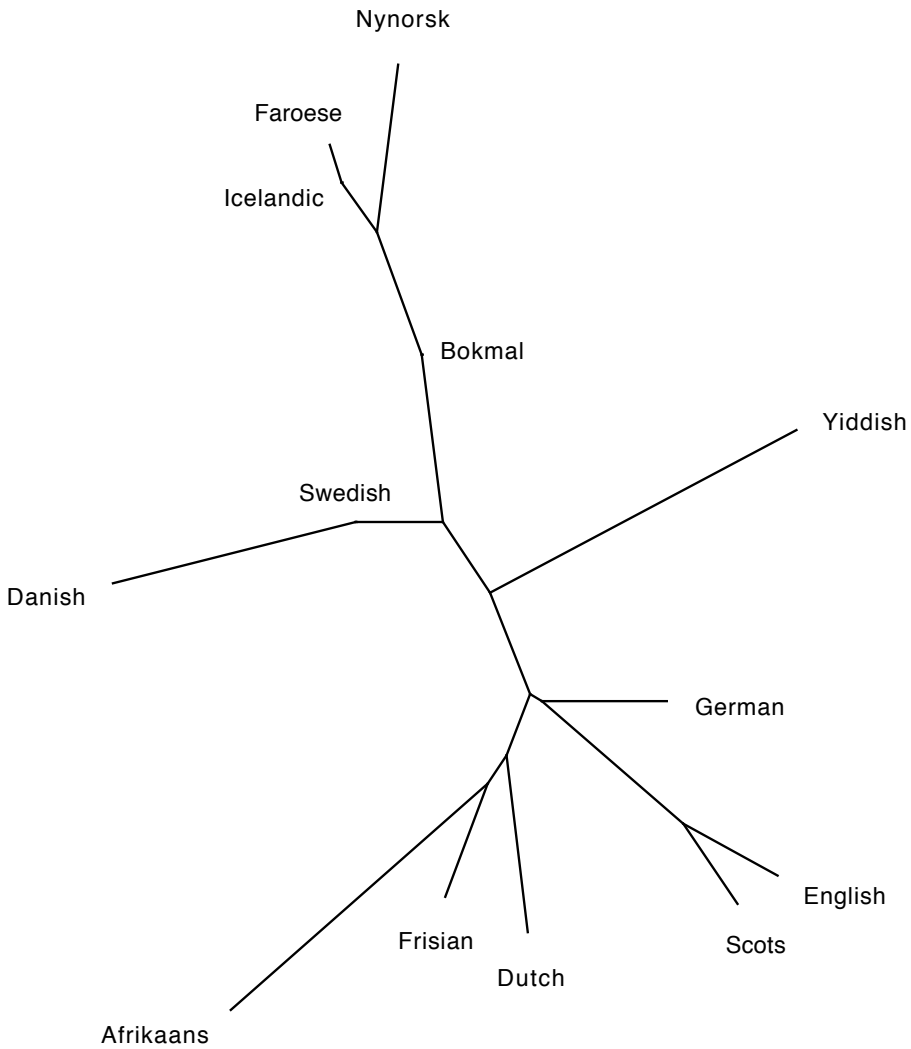
- Linked vector comparison

$$\cos \alpha_{linked}(f, g) = \frac{f^T \cdot \frac{1}{S} \cdot g}{\sqrt{(f^2)^T \cdot \frac{1}{S} \cdot 1_{(g \times f)} \cdot \frac{1}{S} \cdot g^2}}$$

f, g = frequency vectors of contextual words

S = matrix of translational probabilities

Neighbour-joining tree from the linked vector comparison



Conclusions

- Using a distributional definition of meaning is highly useful for language comparison
- Parallel texts are an easy way to obtain comparable contexts across languages
- Differences in the distribution of words in parallel texts is a way to approach differences in meaning of those words
- Meaning differences are phylogenetic informative