Thomas Mayer

Research Unit Quantitative Language Comparison
Ludwig-Maximilians-Universität München
DFG Project "Algorithmic corpus-based approaches to typological comparison"

Algorithmic corpus-based approaches to typological comparison
DFG Project (2011-2014)

The goals of this project are threefold:

First, we will prepare corpora of lesser-studied languages for typological comparisons. Because of the limited amount of research on these languages, these corpora will mainly be unannotated corpora. To be able to investigate unannotated corpora, we will also prepare a **smaller amount of parallel corpora** as a starting point for the automatic analysis.

Second, this project will use existing algorithms and **develop new algorithms to add (approximate) linguistic annotations and extract relevant statistics from the corpora**, allowing for the automatic assessment of typological parameters.

Finally, the main intrinsic goal of this project is to investigate **how much linguistic knowledge of a language is needed to establish a particular typological parameter**.

Parallel corpora have received a lot of attention since the advent of statistical machine translation (Brown et al., 1988) where they serve as training material for the underlying alignment models.

Yet there are only few resources which comprise texts for which translations are available into many different languages. Such texts are here referred to as 'massively parallel texts' (MPT; Cysouw and Wälchli, 2007).

The most well-known MPT is the Bible, which has a long tradition in being used as the basis for language comparison. Apart from that, other religious texts are also available online and can be used as MPTs. One of them is a collection of pamphlets of the Jehova's Witnesses, some of which are available for over 250 languages.

In order to test our methods on a variety of languages, we collected a number of pamphlets from the Watchtower website (`http://www.watchtower.org`) together with their translational equivalents for 146 languages in total (252 question sentences containing a question word in the English version).

We start from a massively parallel text, which we consider as an $n \times m$ matrix consisting of...

n different parallel sentences $S = \{S_1, S_2, S_3, ..., S_n\}$ in m different languages $L = \{L_1, L_2, L_3, ..., L_m\}$.

	Sentence no. 25 (S_{25})	
L_1	why is there a need for a new world	(English, en)
L_2	warum brauchen wir eine neue welt	(German, de)
L_3	защо се нуждаем от нов свят	(Bulgarian, bl)
L_4	por qu se necesita un nuevo mundo	(Spanish, es)
L_5	għala hemm bżonn ta dinja ġdida	(Maltese, mt)
L_6	nukatae míehiã xexeme yeye	(Ewe, ew)
	Sentence no. 93 (S_{93})	
L_1	who will rule with jesus	(English, en)
L_2	wer wird mit jesus regieren	(German, de)
L_3	кой ще управлява с исус	(Bulgarian, bl)
L_4	quiénes gobernarán con jesús	(Spanish, es)
L_5	min se jaħkem ma ġesù	(Maltese, mt)
L_6	amekawoe aɖu fia kple yesu	(Ewe, ew)

SL data-matrix ('sentences × languages')

	L_1	L_2	L_m
S_1	why is it often good to ask questions	warum ist es oft gut fragen zu stellen	...
S_2	why do many stop trying to find answers...	warum hören viele auf nach antworten...	...
S_3	why can we trust that god will undo...	warum können wir uns darauf verlassen...	...
S_4	what does the name jehovah mean	was bedeutet der name jehova	...
S_5	what may we learn about jehovah...	was sagen folgende titel ber jehova...	...
S_6	in what ways is the bible different...	warum ist die bibel ein ganz besonderes...	...
S_7	how can the bible help you cope...	wie kann uns die bibel bei persönlichen...	...
S_8	why can you trust the prophecies...	warum kann man den prophezeiungen...	...
S_9	in what ways is the bible an exciting...	warum kann man sagen dass die bibel...	...
S_{10}	what impresses you about the...	was ist an der verbreitung der bibel...	...
...	...	...	...

each sentence S consists of one or more utterances U:
$S = \{$Why is Jehovah pleased with Abel's gift, and why is he not pleased with Cain's?$\}$
$U_1 = \{$Why is Jehovah pleased with Abel's gift$\}$; $U_2 = \{$and why is he not pleased with Cain's?$\}$

Simplifying assumptions

- most words occur only once per sentence
- no language-specific chunking
- no language-specific recognition of morpheme boundaries (e.g., *question-s*), multi-word expressions (e.g., *por qué*) and phrase structures (e.g., *to ask questions*)

The parallel text can then be encoded as three **sparse matrices**:

UL ('utterances × languages'):	which utterance belongs to which language?
US ('utterances × sentences'):	which utterance belongs to which sentence?
UW ('utterances × words'):	which words occur in which utterance?

UL is defined as...
$\mathbf{UL}_{ij} = 1$ if the utterance i belongs to language j and
$\mathbf{UL}_{ij} = 0$ if not.
Likewise for the other two matrices.

Note the similarity with the wordlist approach where sentences correspond to concepts, utterances to words and words to phonemes/graphemes.

The matrix **US** will be used to compute co-occurrence statistics of all pairs of words, both within and across languages. Basically, we define **O** ('observed co-occurrences') and **E** ('expected co-occurrences') as:

$$\mathbf{O} = \mathbf{WU} \cdot \mathbf{WU^T}$$

$$\mathbf{E} = \mathbf{WU} \cdot \frac{\mathbf{1_{SS}}}{n} \cdot \mathbf{WU^T}$$

The symbol '$\mathbf{1_{ab}}$' refers to a matrix of size $a \times b$ consisting of only 1's

Assuming that the co-occurrence of words follows a poisson process (Quasthoff and Wolff, 2002), the co-occurrence matrix **WW** ('words × words') can be calculated as follows:

$$\mathbf{WW} = -\log[\frac{\mathbf{E^O} \exp(-\mathbf{E})}{\mathbf{O!}}]$$

$$= \mathbf{E} + \log \mathbf{O!} - \mathbf{O} \log \mathbf{E}$$

This **WW** matrix represents a similarity matrix of words based on their co-occurrence in translational equivalents for the respective language pair.

Based on the co-occurrence matrix **WW** we compute concrete alignments (many-to-many mappings between words) for each utterance separately, but for all languages at the same time.

For each utterance U_i we take the subset of the similarity matrix **WW** only including those n words that occur in the row $\mathbf{UW_i}$, i.e., only those words that occur in utterance U_i.

$$WW_i = \begin{pmatrix} ww_{11} & \dots & ww_{1n} \\ \vdots & \vdots & \vdots \\ ww_{n1} & \dots & ww_{nn} \end{pmatrix}$$

We then perform a partitioning on this subset of the similarity matrix **WW** (e.g., affinity propagation clustering; Frey and Dueck, 2007).

The resulting clustering for each sentence identifies groups of words that are similar to each other, which represent words that are to be aligned across languages.

	Sentence no. 93 (S_{93})	
L_1	who will rule with jesus	(English, en)
L_2	wer wird mit jesus regieren	(German, de)
L_3	кой ще управлява с исус	(Bulgarian, bl)
L_4	quiénes gobernarán con jesús	(Spanish, es)
L_5	min se jaħkem ma ġesù	(Maltese, mt)
L_6	amekawoe aḍu fia kple yesu	(Ewe, ew)
...	...	...

With 50 languages as input, the following 10 clusters for those words in the six languages above have been obtained:

1. исус_{bl} jesus_{en} fia_{ew} yesu_{ew} ġesù_{mt} jesús_{es} jesus_{de}
2. кой_{bl} who_{en} min_{mt} wer_{de}
3. $\text{regieren}_{\text{de}}$
4. $\text{управлява}_{\text{bl}}$ aḍu_{ew} $\text{jaħkem}_{\text{mt}}$ $\text{gobernarán}_{\text{es}}$
5. $\text{amekawoe}_{\text{ew}}$ $\text{quiénes}_{\text{es}}$
6. ще_{bl} will_{en} se_{mt} wird_{de}
7. с_{bl} with_{en} con_{es} mit_{de}
8. kple_{ew}
9. ma_{mt}
10. rule_{en}

which yields the following alignment for English, Maltese and Bulgarian:

who_2 will_6 rule_{10} with_7 jesus_1
min_2 se_6 jaħkem_4 ma_7 ġesù_1
кой_2 ще_6 управлява_4 с_7 исус_1

All alignment-clusters from all sentences are summarized as columns in the sparse matrix **WA**, defined as $\mathbf{WA}_{ij} = 1$ when word w_i is part of alignment A_j, and is 0 elsewhere.

WA is then used to derive a similarity between the alignments **AA**. We define both a sparse version of **AA**, based on the number of words that co-occur in a pair of alignments, and a statistical version of **AA**, based on the average similarity between the words in the two alignments:

$$\mathbf{AA}_{sparse} = \mathbf{WA}^{\mathbf{T}} \cdot \mathbf{WA}$$

$$\mathbf{AA}_{statistical} = \frac{\mathbf{WA}^{\mathbf{T}} \cdot \mathbf{WW} \cdot \mathbf{WA}}{\mathbf{WA}^{\mathbf{T}} \cdot 1_{\mathbf{WW}} \cdot \mathbf{WA}}$$

A similarity between languages **LL** can then be defined as:

$$\mathbf{LL} = \mathbf{LA}' \cdot \mathbf{LA}'^{\mathbf{T}}$$

by defining $\mathbf{LA}'$ ('languages $\times$ alignments') as the number of words per language that occur in each selected alignment:

$$\mathbf{LA}' = \mathbf{WL}^{\mathbf{T}} \cdot \mathbf{WA}'$$

As a first step to show that our method yields promising results we ran the method for the 27 Indo-European languages in our sample.

In total, we obtained 6,660 alignments (i.e., 26.4 alignments per sentence on average), with each alignment including on average 9.36 words.

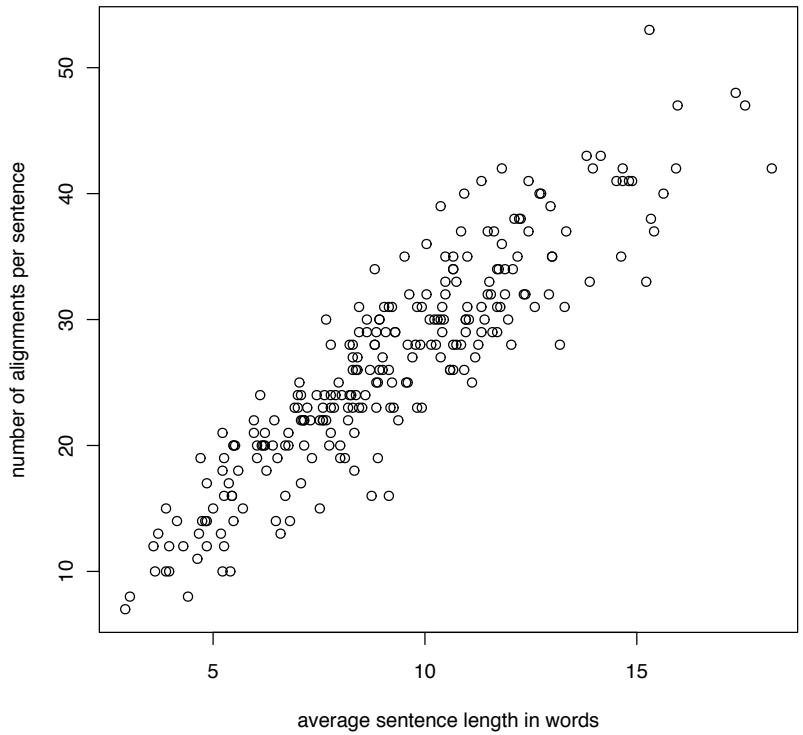

Figure 1: Linear relation (slope of 2.85) between the average number of words per sentence and number of alignments per sentence

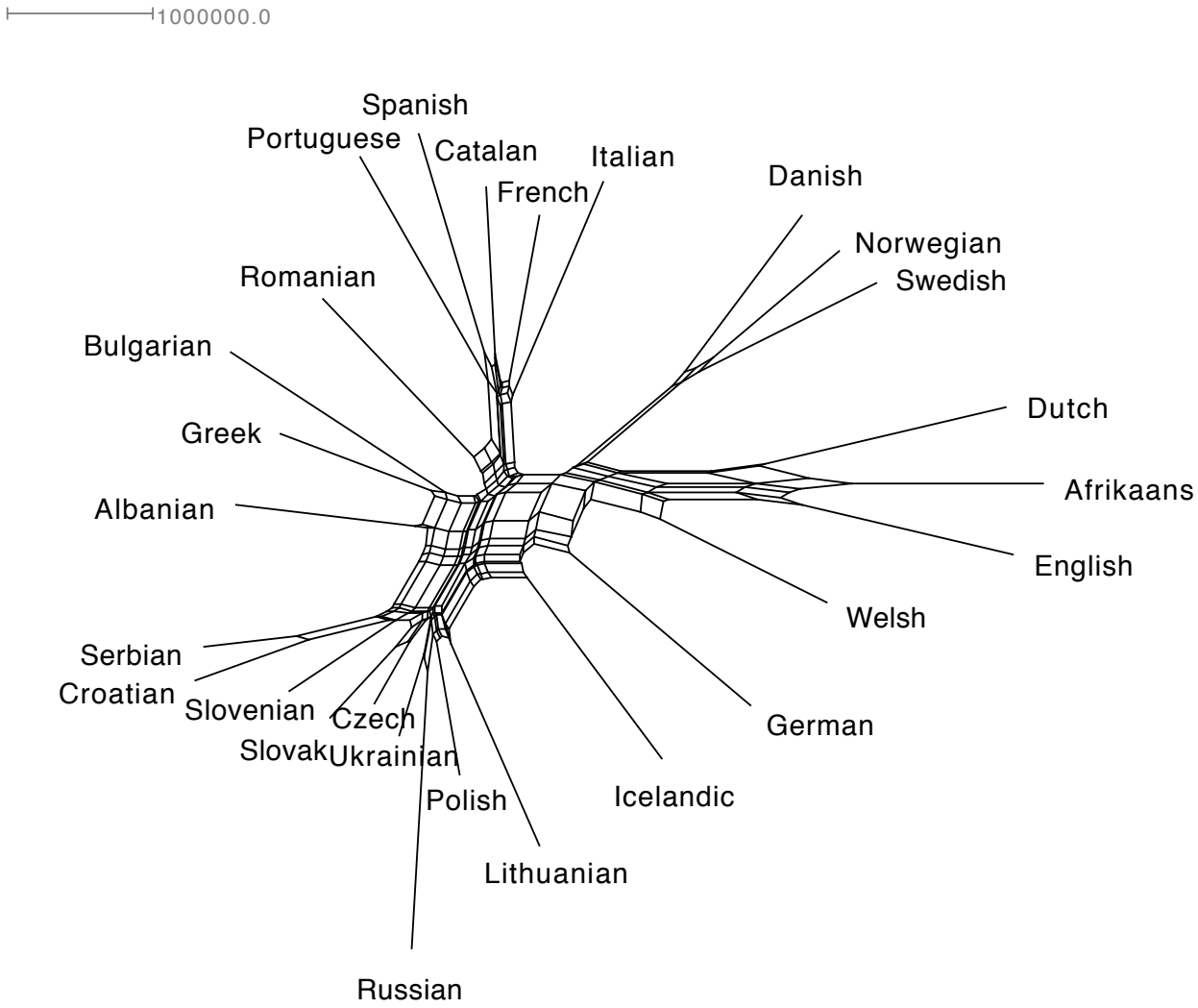

Figure 2: NeighborNet (created with SplitsTree, Huson and Bryant, 2006) of all Indo-European languages in the sample

For this, we selected just the **six** sentences in a sample of 50 languages that were formulated in English with a **who** interrogative, i.e., questions as to the person who did something

1) Who will be resurrected?
2) Who will rule with Jesus?
3) Who created all living things?
4) Who are god's true worshipers on earth today?
5) Who is Jesus Christ?
6) Who is Michael the Archangel?

By using a clustering on the six alignments that comprise English **who**, we ended up having 13 alignments which include words for almost all languages in the six sentences (on average 47.7 words for each sentence).

We computed a language similarity **LL** only on the basis of these 13 alignments, which represents a typology of the structure of PERSON interrogatives.

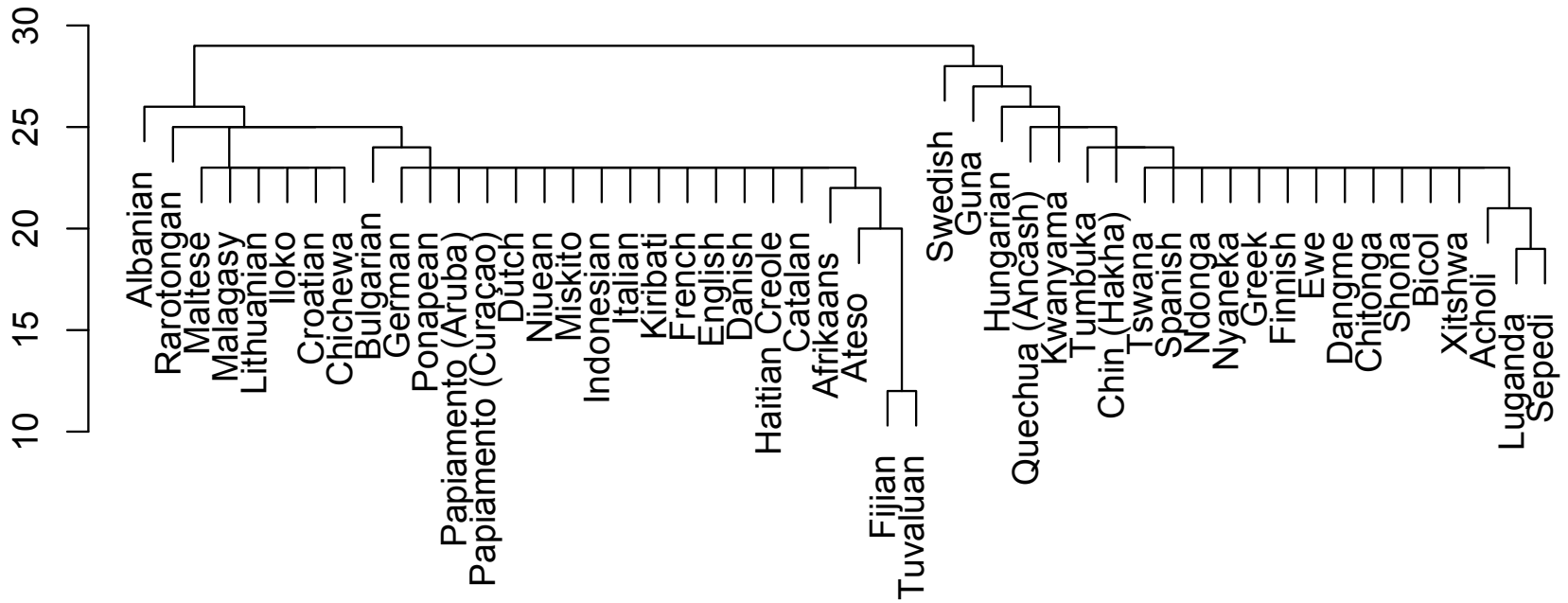

Figure 3: Hierarchical cluster using Ward's minimum variance method (created with R) depicting a typology of languages according to the structure of their PERSON interrogatives

The languages in the right cluster consistently separate the six sentences into two groups:
All languages in the right cluster distinguish between a singular and a plural form of **who**. For example, Finnish uses **ketkä** vs. **kuka** and Spanish **quiénes** vs. **quién**.