Language comparison through sparse multilingual word alignment

Thomas Mayer[1] **Michael Cysouw**[2]

[1]Research Unit Quantitative Language Comparison
Ludwig-Maximilians-Universität München
thommy.mayer@gmail.com

[2] Research Center Deutscher Sprachatlas
Philipps-Universität Marburg
cysouw@uni-marburg.de

EACL Workshop LINGVIS & UNCLH, Avignon, France

Main points of this talk:

- Language comparison:

  we propose a new data source, **parallel texts**

  - historical comparison: as a first step towards a computational approach to Croft's evolutionary theory of language change (where an utterance corresponds to strings of DNA in evolutionary biology)
  - typological comparison:

- Sparse matrices:

  all data structures involved in the calculations are represented as **(sparse) matrices**

- Multilingual word alignment:

  instead of pairwise word alignment we explore the possibilities of the **simultaneous alignment of words in a larger number of languages**

Parallel corpora

- 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).

An evolutionary approach to language change

- So far, phylogenetic methods have been applied using. . .
  - first order: e.g., Swadesh-type lists, non-parallel wordlists
  - second order: e.g., cognate sets, structural characteristics

  . . . data sources for comparison
- We propose yet another first-order data source: **parallel texts**
- Following Croft (2000), we assume that **strings of DNA** in biological evolution correspond to **utterances** in language evolution
- According to this view, **genes** (the functional elements of a string of DNA) correspond to **linguistic structures occurring in utterances** → in this talk we focus on **alignment** as one kind of linguistic structure

utterances vs. words

The choice of translational equivalents in the form of utterances rather than words accounts for the well-known fact that some words cannot be translated accurately between some languages whereas **most utterances in context can be translated accurately**.

Why matrix representations?

- Matrices give a **concise representation of the data** types that we are working with
  → this makes it easier to talk about different types (e.g., SL matrix as a shorthand for the parallel sentences (S) in the various languages (L))
  → this facilitates storing the different types in a pipeline of computational methods
- **Faster computation** with matrix algebra
  → this is especially useful when dealing with large amounts of data. One can fall back on the various methods developed in linear algebra to solve similar problems in an easier way
- The ultimate goal of these representations is that the use of matrix algebra will hint at **decompositions or calculations that are useful for a future analysis** of these data types

Matrix representation

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 qué 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)

⋮

Matrix representation

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*)

Matrix representation

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 **WU** 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:

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

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

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:

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

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

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.

Matrix representation

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)
...	...	...

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

1. исус$_{\text{bl}}$ jesus$_{\text{en}}$ fia$_{\text{ew}}$ yesu$_{\text{ew}}$ ġesù$_{\text{mt}}$ jesús$_{\text{es}}$ jesus$_{\text{de}}$
2. кой$_{\text{bl}}$ who$_{\text{en}}$ min$_{\text{mt}}$ wer$_{\text{de}}$
3. regieren$_{\text{de}}$
4. управлява$_{\text{bl}}$ aɖu$_{\text{ew}}$ jaħkem$_{\text{mt}}$ gobernarán$_{\text{es}}$
5. amekawoe$_{\text{ew}}$ quiénes$_{\text{es}}$
6. ще$_{\text{bl}}$ will$_{\text{en}}$ se$_{\text{mt}}$wird$_{\text{de}}$
7. с$_{\text{bl}}$ with$_{\text{en}}$ con$_{\text{es}}$ mit$_{\text{de}}$
8. kple$_{\text{ew}}$
9. ma$_{\text{mt}}$
10. rule$_{\text{en}}$

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:

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

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

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

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

by defining **LA**′ (‘languages × 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.

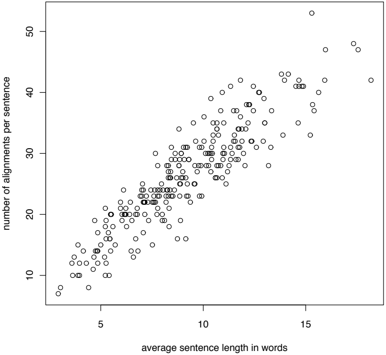

Figure 1: Linear relation (slope of 2.85) between the average number of words per sentence and number of alignments per sentence

Pilot experiment I: Global comparison of Indo-European

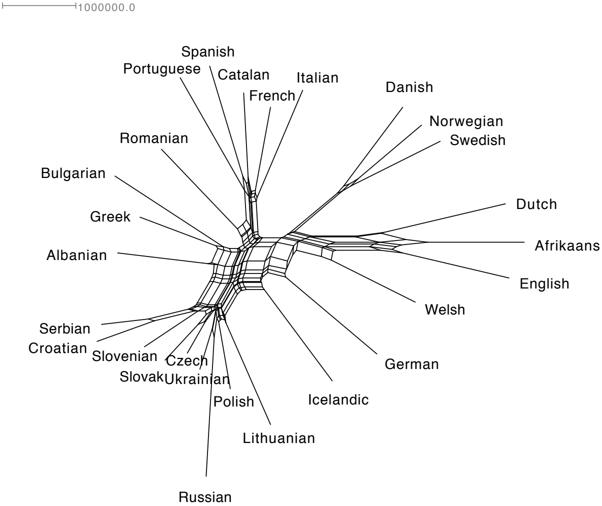

Figure 2: NeighborNet (created with SplitsTree, Huson and Bryant, 2006) of all Indo-European languages in the sample

Conclusions

- The comparison of the IE languages on the basis of their alignment shows an approximate grouping of languages according to the major language families Germanic, Slavic and Romance
- **The form of the words does not play a role** in the comparison, but their frequency of co-occurrence in alignments across languages
- The NeighborNet also exhibits a strong influence of an **areal signal** (Balkan Sprachbund: Albanian, Greek, Bulgarian, Romanian)
  - horizontal transfer due to language contact
  - influence of translationese

  Shared structural features (e.g., the loss of the infinitive, syncretism of dative and genitive case and postposed articles) are particularly prone to lead to a higher similarity in our approach where the alignment of words within sentences is sensitive to the fact that **certain word forms are identical or different even though the exact form of the word is not relevant**

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

S_{79} Who will be resurrected?

S_{93} Who will rule with Jesus?

S_{148} Who created all living things?

S_{176} Who are god's true worshipers on earth today?

S_{245} Who is Jesus Christ?

S_{252} 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.

Pilot experiment II: Typology of PERSON interrogatives

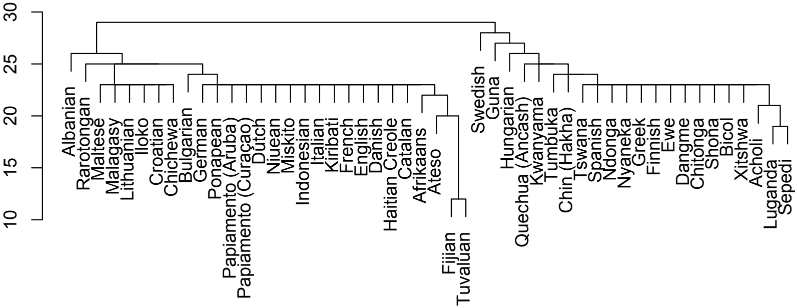

Figure 3: Hierarchical cluster using Ward's minimum variance method (created with R) depicting a typology of languages according to the structure of their PERSON interrogatives

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**.

Pilot experiment II: Typology of PERSON interrogatives

Language	question word
Afrikaans	wie
Albanian*	kush të, cilët
Ateso	ingai/angai
Bulgarian	кой
Catalan	qui
Chichewa	kodi ndani
Croatian	tko
Danish	hvem
Dutch	wie
English	who
French	qui
Fijian	o cei
German	wer
Haitian Creole	kiyès
Iloko	siasino
Indonesian	siapa/siapakah
Italian	chi
Kiribati	antai
Lithuanian	kas
Malagasy	iza
Maltese	min
Miskito	ki
Niuean	ko hai
Papiamento (Curaçao)	ken
Papiamento (Aruba)	ken
Ponapean	ihs
Rarotongan	koai
Tuvaluan	ko oi

(a) First cluster (one question word)

Language	question words
Acholi	anga, angagi
Bicol	siisay, sairisay
Chin (Hakha)	aho/ahodah, ahote
chiTonga	ino nguni, ino mbaani
Dangme	mɛnɔ, mɛnɔmɛ
Ewe	amekae, amekawoe
Finnish	kuka, ketkä
Greek	ποιος, ποιοι
Guna	doa(gi), doamar
Hungarian	ki, kik
Kwanyama	olyelye, oolyelye
Luganda	ani, baani
Ndonga	olye, oolye
Nyaneka	olie, ovalie
Quechua (Ancash)	pitaq, pikunaraq
Sepedi	ke mang, ke bomang
Shona	ndiani, ndivanaani
Spanish	quién, quiénes
Swedish	vem, vilka
Tumbuka	ninjani, mbanjani
Tswana	(ke) mang, ke bomang
xiTshwa	hi, himani

(b) Second cluster (two question words)

Conclusions and future work

Our approach presents a novel method for language comparison:

- proposing a new data source
- looking at structural similarities between languages rather than the forms of words
- considering several languages at the same time for word alignment
- offering a way to represent the various data types and to compute the comparisons with the help of sparse matrices

We consider this as the first step to supplement both the historical and typological comparison of languages. In future work, we plan to. . .

- integrate a more detailed language-specific analysis, like morpheme separation or the recognition of multi-word expressions and phrase structures
- use statistical alignment models (IBM Model 1-3)
- include a validation scheme in order to test how much can be gained from the simultaneous analysis of more than two languages
- refine and formalize the selection of alignments for the comparison of languages, which will enable us to automatically generate typological parameters

References

- Peter F. Brown, John Cocke, Stephen A. Della-Pietra, Vincent J. Della-Pietra, Frederick Jelinek, Robert L. Mercer, and Paul S. Roossin. 1988. A statistical approach to language translation. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, pages 71–76.
- William Croft. 2000. *Explaining Language Change: An Evolutionary Approach*. Harlow: Longman.
- Michael Cysouw and Bernhard Wälchli. 2007. Parallel texts: using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung STUF*, 60(2):95–99.
- Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315:972–976.
- Daniel H. Huson and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267.
- Mark Pagel. 2009. Human language as a culturally transmitted replicator. *Nature Reviews Genetics*, 10:405–415.
- Uwe Quasthoff and Christian Wolff. 2002. The poisson collocation measure and its applications. In *Proceedings of the 2nd International Workshop on Computational Approaches to Collocations*, Vienna, Austria.
- Michel Simard. 1999. Text-translation alignment: Three languages are better than two. In *Proceedings of EMNLP/VLC-99*, pages 2–11.
- Michel Simard. 2000. Text-translation alignment: Aligning three or more versions of a text. In Jean Véronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*, pages 49–67. Dordrecht: Kluwer Academic Publishers.
- Bernhard Wälchli. 2011. Quantifying inner form: A study in morphosemantics. Arbeitspapiere. Bern: Institut für Sprachwissenschaft.