Language comparison through sparse multilingual word alignment

Thomas Mayer
Research Unit
Quantitative Language Comparison
LMU Munich
thommy.mayer@googlemail.com

Michael Cysouw
Research Center
Deutscher Sprachatlas
Philipp University of Marburg
cysouw@uni-marburg.de

Abstract

In this paper, we propose a novel approach to compare languages on the basis of parallel texts. Instead of using word lists or abstract grammatical characteristics to infer (phylogenetic) relationships, we use multilingual alignments of words in sentences to establish measures of language similarity. To this end, we introduce a new method to quickly infer a multilingual alignment of words, using the co-occurrence of words in a massively parallel text (MPT) to simultaneously align a large number of languages. The idea is that a simultaneous multilingual alignment yields a more adequate clustering of words across different languages than the successive analysis of bilingual alignments. Since the method is computationally demanding for a larger number of languages, we reformulate the problem using sparse matrix calculations. The usefulness of the approach is tested on an MPT that has been extracted from pamphlets of the Jehova's Witnesses. Our preliminary experiments show that this approach can supplement both the historical and the typological comparison of languages.

1 Introduction

The application of quantitative methods in historical linguistics has attracted a lot of attention in recent years (cf. Steiner et al. (2011) for a survey). Many ideas have been adapted from evolutionary biology and bioinformatics, where similar problems occur with respect to the genealogical grouping of species and the multiple alignment of strings/sequences. One of the main differences between those areas and attempts to uncover language history is the limited amount of suitable data that can serve as the basis for language comparison. A widely used resource are Swadesh lists or similar collections of translational equivalents in the form of word lists. Likewise, phylogenetic methods have been applied using structural characteristics (e.g., Dunn et al. (2005)). In this paper, we propose yet another data source, namely parallel texts.

Many analogies have been drawn between the evolution of species and languages (see, for instance, Pagel (2009) for such a comparison). One of the central problems is to establish what is the equivalent of the gene in the reproduction of languages. Like in evolutionary biology, where gene sequences in organisms are compared to infer phylogenetic trees, a comparison of the "genes" of language would be most appropriate for a quantitative analysis of languages. Yet, Swadesh-like wordlists or structural characteristics do not neatly fit into this scheme as they are most likely not the basis on which languages are replicated. After all, language is passed on as the expression of propositions, i.e. sentences, which usually consists of more than single words. Hence, following Croft (2000), we assume that the basic unit of replication is a linguistic structure embodied in a concrete utterance.

According to this view, strings of DNA in biological evolution correspond to utterances in language evolution. Accordingly, genes (i.e., the functional elements of a string of DNA) correspond to linguistic structures occurring in those utterances. Linguistic replicators (the "genes" of language) are thus structures in the context of an utterance. Such replicators are not only the words as parts of the sentence but also constructions to express a complex semantic structure, or phonetic

Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 54–62,
Avignon, France, April 23 - 24 2012. ©2012 Association for Computational Linguistics

realizations of a phoneme, to give just a few examples.

In this paper, we want to propose an approach that we consider to be a first step in the direction of using the structure of utterances as the basic unit for the comparison of languages. For this purpose, a multilingual alignment of words in parallel sentences (as the equivalent of utterances in parallel texts) is computed, similar to multi-species alignments of DNA sequences.[1] These alignments are clusters of words from different languages in the parallel translations of the same sentence.[2]

The remainder of the paper is organized as follows. First, we quickly review the position of our approach in relation to the large body of work on parallel text analysis (Section 2). Then we describe the method for the multilingual alignment of words (Section 3). Since the number of languages and sentences that have to be analyzed require a lot of computationally expensive calculations of co-occurrence counts, the whole analysis is reformulated into manipulations of sparse matrices. The various steps are presented in detail to give a better overview of the calculations that are needed to infer the similarities. Subsequently, we give a short description of the material that we used in order to test our method (Section 4). In Section 5 we report on some of the experiments that we carried out, followed by a discussion of the results and their implications. Finally, we conclude with directions for future work in this area.

2 Word Alignment

Alignment of words using parallel texts has been widely applied in the field of statistical machine translation (cf. Koehn (2010)). Alignment methods have largely been employed for bitexts, i.e., parallel texts of two languages (Tiedemann, 2011). In a multilingual context, the same methods could in principle be used for each pair of languages in the sample. One of the goals of this paper, however, is to investigate what can be gained when including additional languages in the alignment process at the same time and not iteratively looking for correspondences in pairs of languages (see Simard (1999), Simard (2000) for a similar approach).

There are basically two approaches to computing word alignments as discussed in the literature (cf. Och and Ney (2003)): (i) statistical alignment models and (ii) heuristic models. The former have traditionally been used for the training of parameters in statistical machine translation and are characterized by their high complexity, which makes them difficult to implement and tune. The latter are considerably simpler and thus easier to implement as they only require a function for the association of words, which is computed from their co-occurrence counts. A wide variety of co-occurrence measures have been employed in the literature. We decided to use a heuristic method for the first steps reported on here, but plan to integrate statistical alignment models for future work.

Using a global co-occurrence measure, we pursue an approach in which the words are compared for each sentence individually, but for all languages at the same time. That is, a co-occurrence matrix is created for each sentence, containing all the words of all languages that occur in the corresponding translational equivalents for that sentence. This matrix then serves as the input for a partitioning algorithm whose results are interpreted as a partial alignment of the sentence. In most cases, the resulting alignments do not include words from all languages. Only those words that are close translational equivalents occur in alignments. This behavior, while not optimal for machine translation, is highly useful for language comparison because differences between languages are implicitly marked as such by splitting different structures into separate alignments.

The languages are then compared on the basis of having words in the same clusters with other languages. The more word forms they share in the same clusters, the more similar the languages are considered to be.[3] The form of the words themselves is thereby of no importance. What counts

[1] The choice of translational equivalents in the form of sentences rather than words accounts for the fact that some words cannot be translated accurately between some languages whereas most sentences can.

[2] In practice, we simply use wordforms as separated by spaces or punctuation instead of any more linguistically sensible notion of 'word'. For better performance, more detailed language-specific analysis is necessary, like morpheme separation, or the recognition of multi-word expressions and phrase structures.

[3] A related approach is discussed in Wälchli (2011). The biggest difference to the present approach is that Wälchli only compares languages pairwise. In addition, he makes use of a global glossing method and not an alignment of words within the same parallel sentence.

is their frequency of co-occurrence in alignments across languages. This is in stark contrast to methods which focus on the form of words with similar meanings (e.g., using Swadesh lists) in order to compute some kind of language similarity. One major disadvantage of the present approach for a comparison of languages from a historical perspective is the fact that such similarities also could be a consequence of language contact. This is a side effect that is shared by the word list approach, in which loanwords have a similar effect on the results. It has to be seen how strongly this influences the final results in order to assess whether our current approach is useful for the quantitative analysis of genealogical relatedness.

3 Method

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 $L = \{L_1, L_2, L_3, ..., L_m\}$. This data-matrix is called $\mathbf{SL}$ ('sentences × languages'). We assume here that the parallel sentences are short enough so that most words occur only once per sentence. Because of this assumption we can ignore the problem of decoding the correct alignment of multiple occurring words, a problem we leave to be tackled in future research. We also ignore the complications of language-specific chunking and simply take spaces and punctuation marks to provide a word-based separation of the sentences into parts. In future research we are planning to include the (language-specific) recognition of bound morphemes, multi-word expressions and phrase structures to allow for more precise cross-language alignment.

Based on these assumptions, we decompose the $\mathbf{SL}$ matrix into two sparse matrices $\mathbf{WS}$ ('words × sentences') and $\mathbf{WL}$ ('words × languages') based on all words w that occur across all languages in the parallel texts. We define them as follows. First, $\mathbf{WS}_{ij} = 1$ when word w_i occurs in sentence S_j, and is 0 elsewhere. Second, $\mathbf{WL}_{ij} = 1$ when word w_i is a word of language L_j, and is 0 elsewhere. The product $\mathbf{WS^T} \cdot \mathbf{WL}$ then results in a matrix of the same size as $\mathbf{SL}$, listing in each cell the number of different words in each sentence. Instead of the current approach of using $\mathbf{WS}$ only for marking the occurrence of a word in a sentence (i.e., a 'bag of words' approach), it is also possible to include the order of words in the sentences by defining $\mathbf{WS}_{ij} = k$ when word w_i occurs in position k in sentence S_j. We will not use this extension in this paper.

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

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

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

$\mathbf{E}_{ij}$ thereby gives the expected number of sentences where w_i and w_j occur in the corresponding translational equivalents, on the assumption that words from different languages are statistically independent of each other and occur at random in the translational equivalents. Note that the symbol '$\mathbf{1_{ab}}$' in our matrix multiplications refers to a matrix of size $a \times b$ consisting of only 1's. Widespread co-occurrence measures are pointwise mutual information, which under these definitions simply is $\log \mathbf{E} - \log \mathbf{O}$, or the cosine similarity, which would be $\frac{\mathbf{O}}{\sqrt{\mathbf{n \cdot E}}}$. However, we assume that the co-occurrence of words follow a poisson process (Quasthoff and Wolff, 2002), which leads us to define the co-occurrence matrix $\mathbf{WW}$ ('words × words') using a poisson distribution as:

$$\begin{aligned}\mathbf{WW} &= -\log[\frac{\mathbf{E^O} \exp(-\mathbf{E})}{\mathbf{O}!}] \\ &= \mathbf{E} + \log \mathbf{O}! - \mathbf{O} \log \mathbf{E}\end{aligned}$$

This $\mathbf{WW}$ matrix represents a similarity matrix of words based on their co-occurrence in translational equivalents for the respective language pair. Using the alignment clustering that is based on the $\mathbf{WW}$ matrices for each sentence, we then decompose the words-by-sentences matrix $\mathbf{WS}$ into two sparse matrices $\mathbf{WA}$ ('words × alignments') and $\mathbf{AS}$ ('alignments × sentences') such that $\mathbf{WS} = \mathbf{WA} \cdot \mathbf{AS}$. This decomposition is the basic innovation of the current paper.

The idea is to compute concrete alignments from the statistical alignments in $\mathbf{WW}$ for each sentence separately, but for all languages at the same time. For each sentence S_i we take the subset of the similarity matrix $\mathbf{WW}$ only including those words that occur in the column $\mathbf{WS_i}$, i.e., only those words that occur in sentence S_i. We then perform a partitioning on this subset of the similarity matrix $\mathbf{WW}$. In this paper we use the affinity propagation clustering approach from Frey and Dueck (2007) to identify the clusters, but this is mainly a practical choice and other methods could be used here as well. The reason for this choice is that this clustering does not require a pre-defined number of clusters, but establishes the optimal number of clusters together with the clustering itself.[4] In addition, it yields an exemplar for each cluster, which is the most typical member of the cluster. This enables an inspection of intermediate results of what the clusters actually contain. 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. Note that we do not force such clusters to include words from all languages, nor do we force any restrictions on the number of words per language in each cluster.[5] In practice, most alignments only include words from a small number of the languages included.

To give a concrete example for the clustering results, consider the English sentence given below (no. 93 in our corpus, see next section) together with its translational equivalents in German, Bulgarian, Spanish, Maltese and Ewe (without punctuation and capitalization).

i. who will rule with jesus (English, en)
ii. wer wird mit jesus regieren (German, de)
iii. кой ще управлява с исус (Bulgarian, bl)
iv. quiénes gobernarán con jesús (Spanish, es)
v. min se jaħkem ma ġesù (Maltese, mt)
vi. amekawoe aḏu fia kple yesu (Ewe, ew)

These six languages are only a subset of the 50 languages that served as input for the matrix $\mathbf{WW}$ where all words that occur in the respective sentence for all 50 languages are listed together with their co-occurrence significance. When restricting the output of the clustering to those words that occur in the six languages given above, however, the following clustering result is 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$_{\text{bl}}$ aḏu$_{\text{ew}}$ 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$_{\text{en}}$

First note that the algorithm does not require all languages to be given in the same script. Bulgarian исус is grouped together with its translational equivalents in cluster 1 even though it does not share any grapheme with them. Rather, words from different languages end up in the same cluster if they behave similarly across languages in terms of their co-occurrence frequency. Further, note that the "question word" clusters 2 and 5 differ in their behavior as will be discussed in more detail in Section 5.2. Also note that the English "rule" and German "regieren" are not included in the cluster 4 with similar translations in the other languages. This turns out to be a side effect of the very low frequency of these words in the current corpus.

In the following, we will refer to these clusters of words as alignments (many-to-many mappings between words) within the same sentence across languages. For instance, sentences i., iii. and v. above would have the following alignment, where indices mark those words that are aligned by the alignment clusters (1.-10.) above:

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 $\mathbf{WA}$, defined as $\mathbf{WA}_{ij} = 1$ when word w_i is part of alignment A_j, and is 0 elsewhere.[6] We also establish the 'book-keeping' matrix $\mathbf{AS}$ to keep track

[4] Instead of a prespecified number of clusters, affinity propagation in fact takes a real number as input for each data point where data points with larger values are more likely to be chosen as exemplars. If no input preference is given for each data point, as we did in our experiments, exemplar preferences are initialized as the median of non infinity values in the input matrix.

[5] Again, this takes into account that some words cannot be translated accurately between some languages.

[6] For instance, the alignment in 2. above contains the four words {кой, who, min, wer}, which are thus marked with 1 whereas all other words have 0 in this column of the $\mathbf{WA}$ matrix.

of which alignment belongs to which sentence, defined as $\mathbf{AS}_{ij} = 1$ when the alignment A_i occurs in sentence S_j, and as 0 elsewhere. The alignment matrix $\mathbf{WA}$ is the basic information to be used for language comparison. For example, the product $\mathbf{WA} \cdot \mathbf{WA}^\mathbf{T}$ represents a sparse version of the words × words similarity matrix $\mathbf{WW}$.

A more interesting usage of $\mathbf{WA}$ is to derive a similarity between the alignments $\mathbf{AA}$. We define both a sparse version of $\mathbf{AA}$, based on the number of words that co-occur in a pair of alignments, and a statistical version of $\mathbf{AA}$, based on the average similarity between the words in the two alignments:

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

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

The $\mathbf{AA}$ matrices will be used to select suitable alignments from the parallel texts to be used for language comparison. Basically, the statistical $\mathbf{AA}$ will be used to identify similar alignments within a single sentence and the sparse $\mathbf{AA}$ will be used to identify similar alignments across different sentences. Using a suitable selection of alignments (we here use the notation $\mathbf{A}'$ for a selection of alignments[7]), a similarity between languages $\mathbf{LL}$ can be defined as:

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

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

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

The similarity between two languages $\mathbf{LL}$ is then basically defined as the number of times words are attested in the selected alignments for both languages. It thus gives an overview of how structurally similar two languages are, where languages are considered to have a more similar structure the more words they share in the alignment clusters.

4 Data

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. For this reason, the last two decades have seen an increasing interest in the collection of parallel corpora for a number of language pairs (Hansard[8]), also including text corpora which contain texts in three or more languages (OPUS[9], Europarl[10], Multext-East[11]). 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; cf. 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. The texts needed some preprocessing to remove HTML markup, and they were aligned with respect to the paragraphs according to the HTML markup. We extracted all paragraphs which consisted of only one sentence in the English version and contained exactly one English question word (*how, who, where, what, why, whom, whose, when, which*) and a question mark at the end. From these we manually excluded all sentences where the "question word" is used with a different function (e.g., where *who* is a relative pronoun rather than a question word). In the end we were left with 252 questions in the English version and the corresponding sentences in the 145 other languages. Note that an English interrogative sentence is not necessarily translated as a question in each other language (e.g., the English question *what is the truth about God?* is simply translated into German as *die Wahrheit über Gott* 'the truth

[7] Note that the prime in this case does not stand for the transpose of a matrix, as it is sometimes used.

[8] `http://www.isi.edu/natural-language/download/hansard/`

[9] `http://opus.lingfil.uu.se`

[10] `http://www.statmt.org/europarl/`

[11] `http://nl.ijs.si/ME/`

about God'). However, such translations appear to be exceptions.

5 Experiments

5.1 Global comparison of Indo-European

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 order to see what kind of global language similarity arises when using the present approach. In our procedure, each sentence is separated into various multilingual alignments. Because the structures of languages are different, not each alignment will span across all languages. Most alignments will be 'sparse', i.e., they will only include words from a subset of all languages included. 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. The number of alignments per sentence turns out to be linearly related to the average number of words per sentence, as shown in Fig. 1. A linear interpolation results in a slope of 2.85, i.e., there are about three times as many alignments per sentence as the average number of words. We expect that this slope depends on the number of languages that are included in the analysis: the more languages, the steeper the slope.

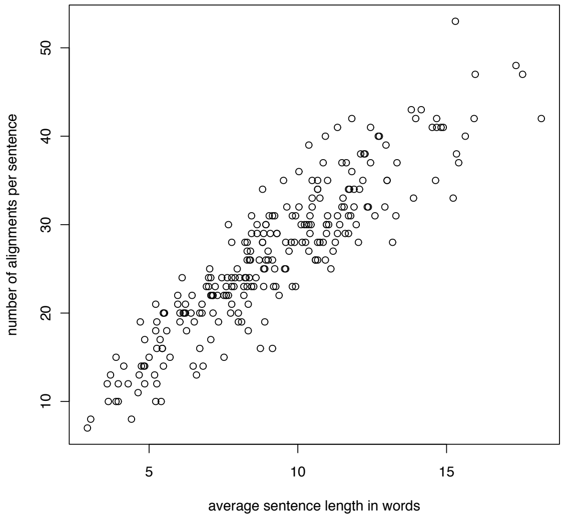

Figure 1: Linear relation between the average number of words per sentence and number of alignments per sentence

We use the $\mathbf{LL}$ matrix as the similarity matrix for languages including all $6,660$ alignments. For each language pair this matrix contains the number of times words from both languages are attested in the same alignment. This similarity matrix is converted into a distance matrix by subtracting the similarity value from the highest value that occurs in the matrix:

$$\mathbf{LL}_{dist} = max(\mathbf{LL}) - \mathbf{LL}$$

This distance matrix $\mathbf{LL_{dist}}$ is transformed into a NeighborNet visualization for an inspection of the structures that are latent in the distance matrix. The NeighborNet in Fig. 2 reveals an approximate grouping of languages according to the major language families, the Germanic family on the right, the Romance family on the top and the Slavic family at the bottom. Note that the sole Celtic language in our sample, Welsh, is included inside the Germanic languages, closest to English. This might be caused by horizontal influence from English on Welsh. Further, the only Baltic language in our sample, Lithuanian, is grouped with the Slavic languages (which is phylogenetically expected behavior in line with Gray and Atkinson (2003)), though note that it is grouped particularly close to Russian and Polish, which suggests more recent horizontal transfer. Interestingly, the separate languages Albanian and Greek roughly group together with two languages from the other families: Romanian (Romance) and Bulgarian (Slavic). This result is not in line with their phylogenetic relatedness but rather reflects a contact situation in which all four languages are part of the Balkan Sprachbund.

Although the NeighborNet visualization exhibits certain outcomes that do not correspond to the attested genealogical relationship of the languages, the method still fares pretty well based on a visual inspection of the resulting NeighborNet. In the divergent cases, the groupings can be explained by the fact that the languages are influenced by the surrounding languages (as is most clear for the Balkan languages) through direct language contact. As mentioned before, a similar problem also exists when using word lists to infer phylogenetic trees when loanwords introduce noise into the calculations and thus lead to a closer relationship of languages than is genealogically tenable. However, in the case of our alignments

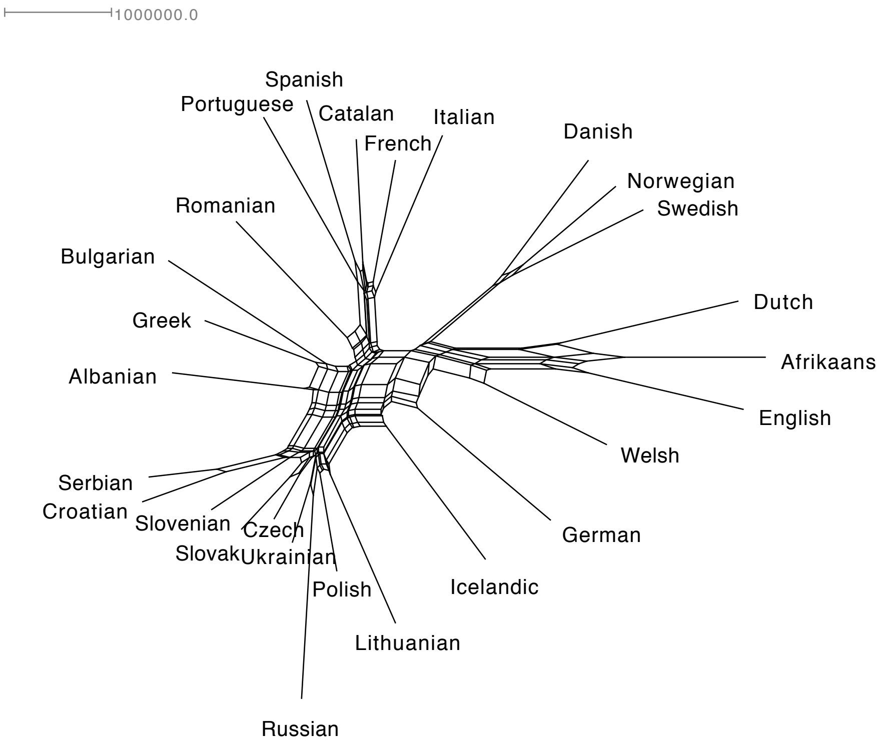

Figure 2: NeighborNet (created with SplitsTree, Huson and Bryant (2006)) of all Indo-European languages in the sample

the influence of language contact is not related to loanwords but to the borrowing of similar constructions or structural features. In the Balkan case, linguists have noted over one hundred such shared structural features, among them the loss of the infinitive, syncretism of dative and genitive case and postposed articles (cf. Joseph (1992) and references therein). These features 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.

5.2 Typology of PERSON interrogatives

A second experiment we conducted involved a closer study of just a few questions in the data at hand to obtain a better impression of the results of the alignment procedure. For this experiment, we took the same 252 questions for a worldwide sample of 50 languages. After running the whole procedure, we selected just the six sentences in the sample that were formulated in English with a *who* interrogative, i.e., questions as to the person who did something. The English sentences are the following:

I Who will be resurrected?
II Who will rule with Jesus?
III Who created all living things?
IV Who are god's true worshipers on earth today?
V Who is Jesus Christ?
VI Who is Michael the Archangel?

We expected to be able to find all translations of English *who* in the alignments. Interestingly, this is not what happened. The six alignments that comprised the English *who* only included words in 23 to 30 other languages in the sample, so we are clearly not finding all translations of *who*. By using a clustering on $\mathbf{AA}_{statistical}$ we were able to find seven more alignments that appear to be highly similar to the six alignments including English *who*. Together, these 13 alignments included words for almost all languages in the six sentences (on average 47.7 words for each sentence). We computed a language similarity $\mathbf{LL}$ only on the basis of these 13 alignments, which represents a typology of the structure of PERSON interrogatives. This typology clearly separates into two clusters of languages, two 'types' so to speak, as can be seen in Fig. 3.

Investigating the reason for these two types, it turns out that the languages in the right cluster of Fig. 3 consistently separate the six sentences into two groups. The first, second, and fourth sentence are differently marked than the third, fifth and sixth sentence. For example, Finnish uses *ketkä* vs. *kuka* and Spanish *quiénes* vs. *quién*. These are both oppositions in number, suggesting that all languages in the right cluster of Fig. 3 distinguish between a singular and a plural form of *who*. Interpreting the meaning of the English sentences quoted above, this distinction makes complete sense. The Ewe form *amekawoe* in example *vi.* (see Section 3) contains the plural marker *-wo*, which distinguishes it from the singular form and indeed correctly clusters together with *quiénes* in the alignment cluster 5.

This example shows that it is possible to use parallel texts to derive a typology of languages for a highly specific characteristic.

6 Conclusion and Future Work

One major problem with using our approach for phylogentic reconstruction is the influence of language contact. Traits of the languages which are not inherited from a common proto-language but are transmitted through contact situations lead to noise in the similarity matrix which does not reflect a genealogical signal. However, other methods also suffer from the shortcoming that language contact cannot be automatically subtracted from the comparison of languages without manual input (such as manually created cognate lists). With translational equivalents, a further problem for the present approach is the influence of translationese on the results. If one version in a language is a direct translation of another language, the structural similarity might get a higher score due to the fact that constructions will be literally translated which otherwise would be expressed differently in that language.

The experiments that have been presented in this paper are only a first step. However, we firmly believe that a multilingual alignment of words is more appropriate for a large-scale comparison of languages than an iterative bilingual alignment. Yet so far we do not have the appropriate evaluation method to prove this. We therefore plan to include a validation scheme in order to test how much can be gained from the simultaneous analysis of more than two languages. Apart from this, we intend to improve the alignment method itself by integrating techniques from statistical alignment models, like adding morpheme separation or phrase structures into the analysis.

Another central problem for the further development of this method is the selection of alignments for the language comparison. As our second experiment showed, just starting from a selection of English words will not automatically generate the corresponding words in the other languages. It is possible to use the $\mathbf{AA}$ matrices to search for further similar alignments, but this procedure is not yet formalized enough to automatically produce language classification for selected linguistic domains (like for the PERSON interrogatives in our experiment). When this step is better understood, we will be able to automatically generate typological parameters for a large number of the world's languages, and thus easily produce more data on which to base future language comparison.

Acknowledgements

This work has been funded by the DFG project "Algorithmic corpus-based approaches to typological comparison". We are grateful to four anonymous reviewers for their valuable comments and suggestions.

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.

Michael Dunn, Angela Terrill, Ger Reesink, R. A. Foley, and Steve C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–5, 9.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315:972–976.

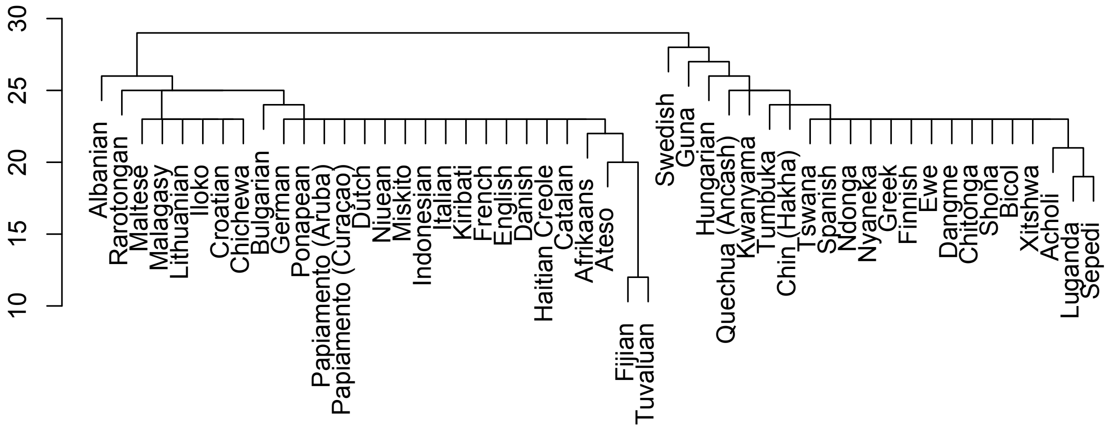

Figure 3: Hierarchical cluster using Ward's minimum variance method (created with R, R Development Core Team (2010)) depicting a typology of languages according to the structure of their PERSON interrogatives

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426:435–439.

Daniel H. Huson and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267.

Brian D. Joseph. 1992. The Balkan languages. In William Bright, editor, *International Encyclopedia of Linguistics*, pages 153–155. Oxford: Oxford University Press.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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.

R Development Core Team, 2010. *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.

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.

Lydia Steiner, Peter F. Stadler, and Michael Cysouw. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1):89–127.

Jörg Tiedemann. 2011. *Bitext Alignment*. Morgan & Claypool Publishers.

Bernhard Wälchli. 2011. Quantifying inner form: A study in morphosemantics. Arbeitspapiere. Bern: Institut für Sprachwissenschaft.