MICHAEL CYSOUW (Leipzig)
CHRISTIAN BIEMANN (Leipzig)
MATTHIAS ONGYERTH (Leipzig)

Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts[1]

We describe a method for the automatic alignment of parallel texts using co-occurrence statistics. The assumption of this approach is that words which are often found together are linked in some way. We employ this assumption to automatically suggest links between words in different languages, using Bible verses as information units. The result is a word-by-word alignment between between different translations of the Bible. The accuracy of our method is evaluated by using Strong's numbers as a benchmark. Overall, the performance is high, indicating that this approach can be used to give an approximate gloss of Bible verses.

1. Introduction

Using parallel texts for linguistic typology is a highly interesting and potentially fruitful approach. However, currently such work is tedious and highly laborious, as every example sentence from every language in the typological sample has to be interpreted individually by a researcher. In this paper, we will propose a method of automatic alignment[2] between translations that could help the interpretation of sentences in a language not intimately known to a researcher, thus possibly speeding up the process of gathering typological data. We envision a system in which a typological researcher selects particular stretches of text from a language of choice because they are considered potentially interesting for a particular linguistic question. Then the system will return the translational equivalents of these sentences in another language, suggesting also an approximate gloss. Of course, the selection, the full analysis, and the interpretation of the sentences will still be left to the typologist.

As an example, consider the verse John 14:6 from the English King James' Version: "Jesus saith unto him: I am the way, the truth, and the life: no man cometh unto the Father, but by me." The Estonian equivalent of this verse is shown in (1) and the Mandarin Chinese equivalent is shown in (2). The glosses given are the glosses suggested by the automatic procedure as described in this paper (unmatched words are indicated by a dash). Although the glosses are not perfect nor complete, they are helpful for a first analysis of these sentences.[3]

[1] We thank BERNHARD WÄLCHLI for useful comments on earlier version of this paper, and we thank BERNHARD COMRIE and GERHARD HEYER for making possible this cooperation between the Max Planck Institute for Evolutionary Anthropology and the University of Leipzig.

[2] Please note that the term 'alignment' is not used here in the linguistic sense (i.e. relating to the marking of arguments), but in the 'normal' meaning of putting things in line.

[3] B. WÄLCHLI (p.c.) informs us that the Estonian gloss does not have any errors. The inclusion of a demoted actor phrase in passive (*minu kaudu*, 'by me') is bad Estonian, but this is a problem with the Bible translation, not with our alignment. H.-J. BIBIKO (p.c.) informs us that the Chinese gloss almost perfect. Only the character glossed as 'but' does not mean *but*.

(1) Estonian (Uralic)

Jeesus	*ütleb*	*temale:*	*Mina*	*olen*	*tee*	*ja*	*tõde*	*ja*	*elu,*
Jesus	saith	him	I	am	way	and	truth	and	life
ükski	*ei*	*saa*	*Isa*	*juure*	*muidu*	*kui*	*Minu*	*kaudu!*	
man	no	–	Father	unto	–	–	I	by	

(2) Mandarin Chinese (Sino-Tibetan)

耶穌	說	我	就	是	道路、	真	理、	生命；			
Jesus	saith	I	–	am	way	truth	–	life			
若	不	藉	著	我，	沒有	人	能	到	父	那	裡	去。
–	but	by	–	I	no	man	–	–	father	–	–	–

This paper is organised as follows. First, there is some general discussion on our approach to automatic alignment. In Section 2, we present a short survey of the problem of automatic word-by-word alignment. In Section 3, the fundamental principle of our approach to this problem is presented, viz. *co-occurrence statistics*, which is based on the assumption that words are linked, when they are often found together in a corpus of a particular language. Then, in Section 4, we discuss how these statistics can be used for alignment between different languages. The basic idea is to count co-occurrences in the same sentences between two difference languages. Such count will be called *trans-co-occurrences*.

The second part of this paper presents an application of this method. Here, we attempt to align different translations of the Bible. In Section 5, we describe how we extracted a sentence-by-sentence alignment from Bible translations, and how we prepared such translations for our analysis. In Section 6, the sentence-by-sentence alignment is turned into a word-by-word alignment using trans-co-occurrences. Finally, in Section 7 the resulting word alignments are evaluated using a concordance-method as used in Bible exegesis: the so-called Strong's Numbers. The results of this evaluation are promising, suggesting that our approach to the alignment of parallel texts is worthwhile, and should be pursued further.

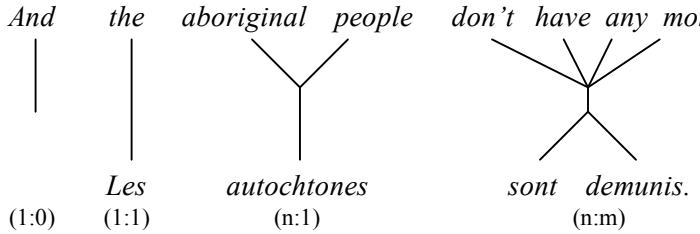
2. Word alignment

The task of word alignment is to link wordforms in a text to its correspondences in the translated text in another language, in such a way that the connected words supply the same contents. Computational proposals for this problem have been made starting in the late 1980's (cf. VÉRONIS 2000 for a survey). For most parallel texts, the problem already starts with the alignment of sentences. Given a text and its integral translation, which sentence in language *B* is to be considered the translation of a sentence in langage *A*? Much of the literature on automatic alignment deals with this problem. However, for our current task of aligning Bible translations, the sentence alignment is to a great extend already provided in the form of verse numbering, which is included in all Bible translations (cf. Section 5 for more details). The task thus is reduced to producing word-by-word linkage on the basis of given sentence-by-sentence (or better verse-by-verse) alignment.

The kinds of linkage attested varies depending on the typological structure of the languages and on the freedom of the translation. An example of word-by-word alignment is presented in Figure 1, following the examples and analysis by BROWN *et al.* (1990; 1993). A commonly attested type of linkage is a *1:1* association, exemplified here with the link between *The* and *Les*. In this case we can assume that the meaning of both wordforms are roughly equal. In *1:0* linkage, the equivalent of a particular wordform is not present in the translation, as shown for *And* in Figure 1. Often, words have to be associated with multiple words in the other language. This are so-called *1:n* or *n:1* associations, regularly found with compounds or fixed constructions (cf. *autochtones* in the figure). Figure 1 also highlights the most complicated case: a general *n:m* alignment, where on both sides multiple words are linked together. Although it is possible to divide these multi-word constructions into smaller parts in both languages separately, this cannot be done simultaneously in both languages in a compatible way. Such general *n:m* alignments will occur with high frequency when two rather strongly agglutinating or polysynthetic languages are aligned.

In this paper, we will approach the problem of word alignment using co-occurrence statistics. This method has, to our knowledge, not been attempted for the alignment of parallel texts. The research reported on here is only a first attempt at using this method for this goal, and there are various improvements possible. However, even with the rather basic implementation used here, we are already getting fairly good results, suggesting that this approach is worthwhile pursuing.

Figure 1. An alignment between an English sentence and a French translational equivalent showing different kinds of linkage.



3. Using co-occurrence statistics

The goal of co-occurrence statistics is to extract pairs of words that are associated from a corpus. The underlying assumption is that while generating text, people are complying to syntactic and semantic restrictions of their (natural) language in order to produce correct sentences. When analyzing a large quantity of text (a text corpus), words that tend to appear together will reflect these linguistic restrictions. While it is generally possible to produce sentences containing arbitrary pairs of

words, in most of the cases the words appearing together will have something to do with each other and statistics will be able cut out the noise.

The joint occurrence of words within a well-defined unit of information, for example the sentence, a whole document, or a word window,[4] is called a co-occurrence. The simplest co-occurrence statistics would be to count how often two words co-occurr within all units of information in the corpus. However, because more frequent words have higher probabilities in appearing together with any word, just because they are frequent, this will not give meaningful associations. Therefore, a significance measure is applied that takes the single word frequencies as well as their joint frequency into account. In our experiments, we use a log-likelihood measure that, intuitively speaking, measures the amount of surprise to see two words co-occurring together as often as they do, compared to the statistical expected number of co-occurrences if we assume independence of occurrence. Here, the significance values for the co-occurrence of two words A and B are calculated according to the formula as shown in (3), cf. BIEMANN *et al.* (2004a).

$$(3) \qquad sig(A,B) = \frac{x - k\log x + \log k!}{\log n}$$

n = number of units of information in the corpus
k = number of joint occurrences of A and B within a unit of information
$x = ab/n$
a = number of occurrences of A in the corpus
b = number of occurrences of B in the corpus

The significances are computed for every pair of words in the corpus. The significance values give us the possibility to rank the co-occurrences of a given word, as higher significance values denote a higher degrees of association. Normally, such statistics are applied on monolingual corpora, and the results are semantic nets. Semantically related words tend to show a high degree of association.[5]

4. Trans-co-occurrences

When applying co-occurrence analysis to multi-lingual parallel texts, we are interested in the association between pairs of wordforms, each from a different language. In that usage, co-occurrence statistics can automatically extract translational equivalents of wordforms, given a sentence-aligned bilingual corpus. Given a sentence translation pair we merely calculate significant co-occurrences between wordforms from different languages and call them *trans-co-occurrences*. If a wordform A in the first language is always translated into wordform B in the second

[4] A *word window* is a stretch of text defined relative to a central word X within a given window size S. The word window around X consists of all words occuring next to X up to maximally S words away. For example, the window of size three around the word 'text' as occurring in the first line of this footnote, consists of the words {a, stretch, of, defined, relative, to}.

[5] This property can be used to create semantic networks for short texts or spoken language streams as discussed in BIEMANN *et al.* (2004b)

language, then *B* will be the highest ranked trans-co-occurrence of *A*. In contrast, often *A* will have various high ranked trans-co-occurrences, normally all with clearly smaller significance values, which represent alternatively possible translations. In this general case, there are several possibilities to translate a wordform from one language into another. In this situation, the most prominent translation will be ranked highest, followed by less prominent translations and finally noise.

Given the data obtained by trans-co-occurrence statistics, it is possible to construct dictionaries from parallel texts in a fully automatic way.[6] All trans-co-occurrences above some significance threshold will be entered in the dictionary. The quality, as compared to manually compiled dictionaries, can be estimated at 60%–80% correctness (SAHLGREN 2004, BIEMANN & QUASTHOFF forthcoming). However, here we are currently not interested in building up dictionaries, including all possible meanings of a particular word, but in word-by-word alignment between two given translational equivalents; in our case of the Bible.

5. Preparations: sentence alignment and markup

For our research, we used Bible translations from the SWORD PROJECT as parallel corpora.[7] To calculate the trans-co-occurrences, two Bibles were merged to a new bilingual bible. Using the Bible's verse numbering as anchors, we combined corresponding sentences to a new longer sentence through concatenation. In principle, we could have simply concatenated whole verses, but we decided to try to restrict the information unit because we were afraid that verses would be too long to yield significant co-occurrences.[8] We tried to restrict the information unit to, roughly, the size of a sentence. To achieve this, we first splitted verses into smaller parts, using full stops and semicolons as separators. If the number of parts obtained is identical for the two languages, then we splitted the verse. However, if the number of parts is not identical, we kept to the complete verse. For example, consider the verse Genesis 1:2 in the English King James Version (KJV) and the German Luther translation as shown in (4).

(4) a. And the earth was without form, and void;
and darkness was upon the face of the deep.
And the Spirit of God moved upon the face of the waters.

b. Und die Erde war wüst und leer, und es war finster auf der Tiefe;
und der Geist Gottes schwebte auf dem Wasser.

---

[6] Note that such an automatically generated dictionary would be a dictionary of wordforms, and not the classical type of linguistic dictionaries only listing lexemes.

[7] http://www.crosswire.org/sword/index.jsp

[8] With hindsight, seeing the results of our investigation, we now think that this step was not necessary. The algorithm that we have used seems to be robust enough to cope with longer information units, like whole verses of the Bible. However, it is to be expected that using larger information units requires more instances (i.e. parallel units) to get reliabale statistics. Of course, by taking larger units, we end up with less units, and would probably get worse results.

As can be seen from this example, after splitting the verse the number of obtained parts differs between the two languages. The English version (4a) consists of three parts, but the German translation (4b) only consists of two parts. So in this case, we are unable to restrain the information unit. The whole verses are simply concatenated into a bilingual sentence, as shown in (5). For the automatic distinction of the languages, each word was marked with language-identifying tags, like '@en' for English or '@de' for German, as shown in (6).

(5) And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters. Und die Erde war wüst und leer, und es war finster auf der Tiefe; und der Geist Gottes schwebte auf dem Wasser.

(6) And@en the@en earth@en was@en without@en form@en and@en void@en and@en darkness@en was@en upon@en the@en face@en of@en the@en deep@en And@en the@en Spirit@en of@en God@en moved@en upon@en the@en face@en of@en the@en waters@en Und@de die@de Erde@de war@de wüst@de und@de leer@de und@de es@de war@de finster@de auf@de der@de Tiefe@de und@de der@de Geist@de Gottes@de schwebte@de auf@de dem@de Wasser@de

Following this approach, two Bible translations can be combined into one language-tagged bilingual Bible. This bilingual text can then be used to compute the trans-co-occurrences for each word.[9]

6. Algorithm for word alignment

Using the trans-co-occurrence statistics, any wordform in a particular sentence from the Bible will now be linked to a wordform in the other language (we used the occurrence of spaces in the text as wordform delimiters). To demonstrate our approach to such word alignment, consider the verse Luke 11:4, as shown in (7) – the English KJV translation in (7a) and the German Luther version in (7b).

(7) a. And lead us not into temptation; but deliver us from evil.
b. Und führe uns nicht in Versuchung, sondern erlöse uns von dem Übel.

From this verse, we have selected the English words *temptation* and *deliver* as exemplars. The German trans-co-occurrences of these English words are tabulated in Table 1 and Table 2, respectively, ordered by the co-occurence significance. The highest ranked words are thus considered to the best overal translational equivalent. However, these tables are based on the whole Bible, so all kind of words do occur, irrespective of the actual words that are found in the German version of the verse Luke 11:4. (The words that occur in this verse are printed in boldface in the tables.) If we would simply take the highest ranked word also present in the Ger-

[9] The procedure to compute the (trans-)co-occernces is described in detail in BIEMANN *et al.* (2004a).

man sentence as the best match, then the English *temptation* is correctly linked to the German *Versuchung*. However, as can be seen from Table 2, the English *deliver* is then wrongly linked to the German *nicht*. The pair (*deliver*, *nicht)* has a higher significance value than the correct pair *(deliver*, *erlöse*). This error sometimes occurs with highly frequent words like *nicht* or *in*.

The basic idea to alleviate this problem is to combine the ranks of the significance statistics looking from English to German with the statistics when looking from German to English. For example, the English *deliver* suggested *nicht* as the best match (on rank 15). However, when we look at the trans-co-occurrence statistics for the German word *nicht*, the English word *deliver* is only ranked as match number 44. In contrast, for the German word *erlöse*, the English word *deliver* ends up as the highest ranked trans-co-occurrant, though it was only ranked on number 19 in Table 2. The pair (*deliver*, *nicht*) has thus ranks 15 and 44, which seems intuitively worse than the pair (*deliver*, *erlöse*) with ranks 19 and 1. We formalized this intuition by defining a MATCH VALUE *m* for a pair of English-German words as shown in (8), based on the multiplication of the two rank-numbers.[10] On the basis of this value we get the right match, because the match value *m*(*deliver*, *erlöse*) is 0.229, which is clearly higher than the match value *m*(*deliver*, *nicht*), which is 0.039.

$$(8) \qquad m(e,g) = \frac{1}{\sqrt{rank_e(g) \cdot rank_g(e)}}$$

Table 1. Ranked German trans-co-occurrences of the English word *temptation*. A selection of words from the German version of Luke 11:4 are printed in boldface.

rank	word	overall corpus frequency	number of co-occurrences	co-occurrence significance
1	**Versuchung**	10	9	59
2	fallet	6	4	26
3	Anfechtung	8	4	25
4	verstocket	4	2	13
5	betet	39	3	13
…				
7	**erlöse**	12	2	11
10	**Übel**	61	2	8
12	**nicht**	7541	11	7

[10] The square root in this formula prevents the *m* values from becoming small very quickly, which might lead to many, possibly confusing, decimal zeros. However, this use of the square root is basically irrelevant, as we are only interested at the relative ordering of the resulting *m* values, and not at their absolute magnitude.

Table 2. Ranked German trans-co-occurrences of the English word *deliver*. A selection of words from the German version of Luke 11:4 are printed in boldface.

rank	word	overall corpus frequency	number of co-occurrences	co-occurrence significance
1	erretten	79	71	260
2	errette	37	34	126
3	Hand	1052	79	109
4	Hände	408	45	78
5	geben	592	47	68
…				
15	**nicht**	7541	117	27
19	**erlöse**	12	7	24
22	**uns**	1525	39	22
59	**führe**	42	5	10
70	**Versuchung**	10	3	9

In this way, the best translational equivalent for a particular word can be found with rather great precision (see the next section for an evaluation of this approach).[11] However, the match value is even more informative because the height gives an indication of how good is the best match that is found. The best possible result is achieved when the matched words are both the higest ranked trans-co-occurrences. Both ranks are then one, and the resulting match value *m* is 1.00. If the matched pair is less directly equivalent, the match value will be lower (cf. *m*(*deliver*, *erlöse*) = 0.229 as discussed above). The height of this value can be used to select only the best translations. Allowing also lower valued matches, more words are actually linked to a translation. However, there will also be some more errors included. This trade-off is investigated in the next section.

7. Using Strong's Numbers as a benchmark

To evaluate the results of our algorithm, we used the so-called 'Strong's Numbers' that are available for some Bible translations. These numbers are annotations added to a Bible text following a system devised by JAMES STRONG in the 19th century. JAMES STRONG (1822-1894) was professor of exegetical theology at Drew Theological Seminary (Madison, New Jersey). Under his guidance, an exhaustive concordance between the King James Version (KJV) of the Bible and the Hebrew

[11] The resulting tables of trans-co-occurrences are a higly valuable resource for other research as well. Note, for example, that it is also possible to use the trans-co-occurrence statistics, as obtained by analysis of the Bible, for the translation of other, yet untranslated texts. However, we can not use the bidirectional match value in that case, but only the ranking as implicit in the trans-co-occurrence statistics.

Old Testament (i.e. the Masoretic Text, called *Tanakh* in Hebrew) and the Greek New Testament (i.e. the *Textus Receptus*) was compiled, apparently with the help of more than a hundred unnamed colleagues. This concordance first appeared in 1890. It is based on a dictionary of all words occurring in the Hebrew and Greek Bibles, which are numbered along their alphabetical order. These numbers are then inserted in the English text of the KJV. Following this example, the same numbers were later also added to various other translations of the Bible.

As an example, consider the verse Revelation of John 1:8 from the New Testament in the KJV translation, as shown in (9). The Greek letter *A*, translated into English as 'Alpha', is the first entry in the Greek alphabetical listing. Accordingly, the word 'Alpha' in the KJV translation is marked with the number <1> behind it. The main difference between these Strong's Numbers and a modern XML-style mark-up is that the Strong Numbers only mark the end of the entry and not the start. This leads to some problems for automatic processing, because it is not clear exactly to which part a Strong's Number refers. For example, the words *is to come* in (9) are not individually marked by a Strong's Number, but only as a group. In most cases, the Strong's Number appears to be placed immediately following the main lexical equivalent of the word in the Greek or Hebrew text. We decided to include only this last word before a Strong's Number for the evaluation of our algorithm. Also note that in some cases there are multiple Strong's Numbers associated with one part of the English translation (e.g. the same phrase *is to come*, associated with the numbers 2064 and 3801). This situation arises because in some cases there are multiple words in the Greek or Hebrew texts which are translated as just one word or phrase into English. We included both numbers for testing the results of our algorithm.

(9) I <1473> am < 1510> **Alpha <1>** and <2532> Omega <5598>, the beginning <746> and <2532> the ending <5056>, saith <3004> the Lord <2962>, which <3588> is <5607, 3801>, and <2532> which <3588> was <2258, 3801>, and <2532> which <3588> **is to come <2064, 3801>**, the Almighty <3841>. (KJV, Rev. 1:8)

When two translations of the Bible are both marked with Strong's Numbers, then these numbers can be used to evaluate an automatically generated alignment. There are four different situations that can occur when comparing the automatic alignment with the Strong's Numbers:

- **Correct:** the aligned words are both followed by a Strong's Number, and these numbers are identical (in case there is only one number) or show an overlap (in case there are multiple numbers)
- **Error:** the aligned words are both followed by a Strong's Numbers, but these numbers are different (in case there is only one number) or do not show any overlap (in case there are multiple numbers)
- **One-sided miss:** only one of the aligned words is followed by a Strong's Number, but the other is not.
- **Uninformative:** both aligned words are not followed by a Strong's Number.

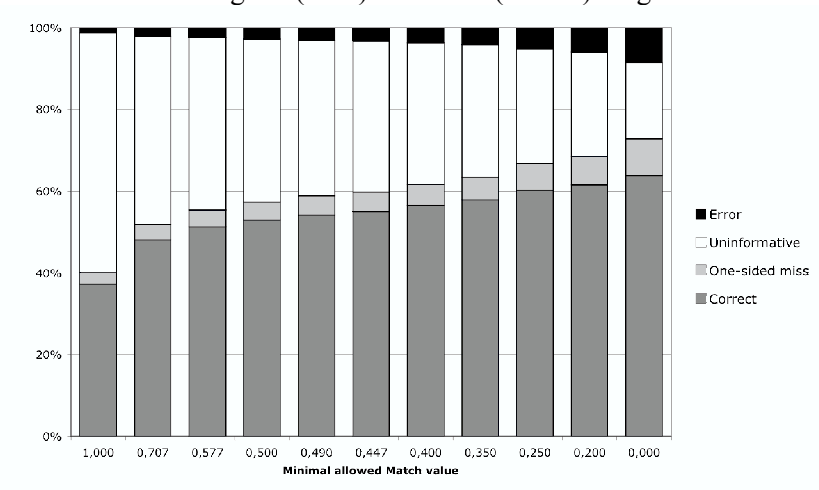
As an example, compare the KJV translation in (9) with the German translation by Luther in (10). When the automatic alignment aligns the English *I* to the German *Ich*, this is counted as correct because both are followed by the same Strong's Number <1473>. However, if the algorithm aligns *am* with *ich*, this would clearly be an error, as the words are followed by different Strong's Numbers. One-sided misses occur for example when the English *Lord* is (correctly) aligned with *Gott*. Although this is correct, this alignment cannot be validated because there is no Strong's Number directly following the German *Gott*. From some random inspection of such cases, we suspect that the far majority of such one-sided misses are actually correct alignments that are obscured by the specific placement of the Strong's Numbers in the text. Finally, there are alignments that cannot be interpreted because both words are not followed by a Strong's Number. For example, neither the article *the* nor *der* are followed by a Strong's Number, and are thus uninformative for the evaluation.

(10) Ich <1473> bin <1510> das A <1> und <2532> das O <5598> , der Anfang <746> und <2532> das Ende <5056> , spricht <3004> Gott der HERR <2962> , der <3588> da ist <3801> und <2532> der <3588> da war <2258> <3801> und <2532> der <3588> da kommt <2064> <3801> , der Allmächtige <3841> . (Luther, Rev. 1:8)

The actual number of errors and correct alignments depends on the MATCH VALUE *m(e,g)*, as defined in the previous section. The match value gives an indication how good the algorithm evaluates a particular alignment of two words between the translations. An alignment with the highest possible match value of 1.00 means that the algorithm rates this as a good match; a lower match values indicates less confidence. In Figure 2, we show the evaluation of the English (KJV) - German (Luther) alignment, depending on the allowed match values. In the first column, only the alignments with a match value of 1.00 are shown. As can be seen in Figure 2, more than 50% of these alignments are uninformative. If lower match values are allowed, this portion becomes smaller, but also the number of errors increases. To show this trade-off between accuracy and overall performance, we defined measures for precision and recall on the basis of these validations, as shown in (11). These values for precision and recall are rather conservative and thus very probably lower than the actual performance of the automatic alignment. We expect that most of the one-sided misses and many of the uninformative cases are actually correct alignments. However, we have no way to assess that more precisely at this point.

(11) Precision = correct / correct + error + one-sided miss
Recall = correct / all alignments

Figure 2. Evaluation of English (KJV) - German (Luther) Alignment



We computed the precision and recall for every match value (i.e. for every column in Figure 2). The resulting values are plotted in Figure 3, connected by a line There are two lines shown in this figure because we performed the alignment directionally. One line in the figure represents the precision and recall for the direction where we started with the English translation and then tried to find the best match in the German translation. The other line represents the inverted procedure. Interestingly, the precision from English to German is better than the other way around, although the recall roughly remains the same. This is probably caused by the fact that German has more mophology than English, and consequently the German translation has less words. The resulting major difference is that the number of one-sided misses is clearly higher for the direction German to English.

Figure 3. Trade-off between precision and recall for the English-German alignment

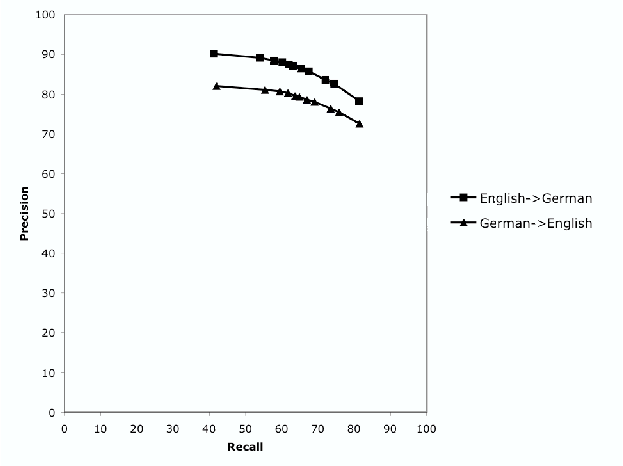
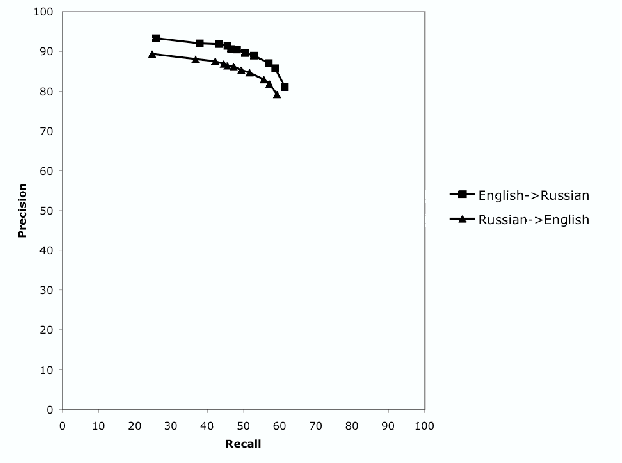


Figure 4. Trade-off between precision and recall for the English-Russian alignment

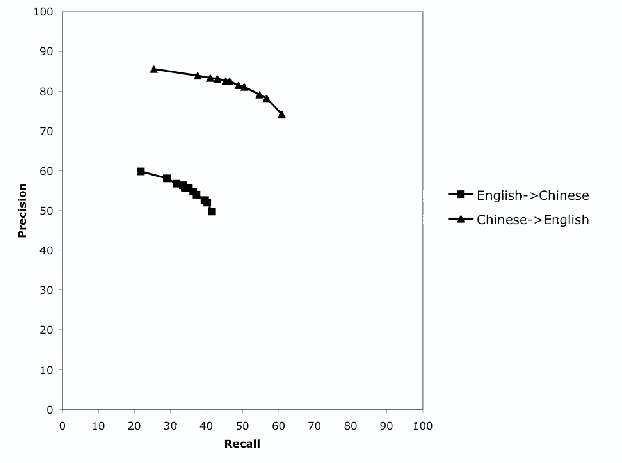


We performed the same evaluation for the alignment of the English KJV translation with the Russian 'Synodal' translation from 1876. The results of the evaluation using the Strong's Numbers is shown in Figure 4. The precision is comparable to the English-German alignment, but the recall is much worse. This is the result of a much higher fraction of uninformative alignments. This is probably caused by the multitude of textual variants of the 'original' texts. The Synodal text of the Russian orthodox Church is probably based on a different original as the English KJV translation (see DE VRIES, this issue).

Finally, we also evaluated an English-Chinese automatic alignment by using the Chinese 'Traditional Union' translation, which has also been annotated by Strong's Numbers. The results are shown in Figure 5. The first aspect to take note of is the large discrepancy between the two directions of alignment. The alignment from English to Chinese is much worse than the alignment from Chinese to English, although for the alignments with German and Russian this direction even performed slightly better. The reason for this large discrepancy is that we did not parse the Chinese text for words.[12] The algorithm simply looked for the best match between any Chinese character and any word in the English text. However, most lexical words in the English text is translated by multiple Chinese characters. Now, for the evaluation of our algorithm we also took the first Chinese character before any Strong's Number. If we start from an English word followed by a Strong's Number, the best match will very often not be the Chinese character directly in front of the Strong's Number, but one of the other characters that also are part of the translation. As a result, we get a very high proportion of one-sided misses for the direction English to Chinese, which diminishes the precision. In contrast, for the direction from Chinese to English, the precision is roughly on the same level as for the alignment from German to English. The recall is worse because of a much higher proportion of uninformative matches.

[12] Of course, this could have been done, e.g. by http://www.mandarintools.com/segmenter.html.

Figure 5. Trade-off between precision and recall for the English-Chinese alignment



This directional difference with the Chinese-English alignment suggests an interesting consequence for the alignment between English and morphologically more complex languages (and that is why we did not parse the Chinese text for words). English could be considered a much more synthetic language compared to the Chinese *script*, as most English words map onto multiple Chinese characters. Of course, such a comparison does not make any sense linguistically. However, this way to look at it argues that the results from our alignment between English words and Chinese characters might be interpreted as showing what would happen if we would try to align English to a more synthetic language. Starting from the morphologically more complex language is difficult for our algorithm (cf. the direction English to Chinese). However, using the alignment from the more isolating language to the more synthetic language seems to give relatively good results (cf. from Chinese to English), even though the structure of the languages are very different. Of course, it would be better to check this claim by actually trying to align the English text to a language with a more complex morphology. Our algorithm does have no problem providing an alignment between English and, say, the Swahili New Testament (which is also available electronically as open source), but we have no way to automatically check such an alignment because there are no Strong's Number added to the Swahili translation (nor to any other translations of morphologically more complex languages).

8. Conclusions

The usage of trans-co-occurrences is a highly promising method to establish translational equivalents in parallel texts. Even in the simple and straighforward version that we used in this paper, the results are already fairly good. At least good enough to provide typologists with an approximate gloss of a stretch of text, which can then subsequently be analysed in more detail by hand.

An important characteristic of our algorithm, which makes it even more interesting for typology, is that there is no knowledge needed about the languages that are to be combined. The algorithm is completely language-independent. The only information that is assumed is an aligned information unit (in our case, the Bible verses) and a word-separator (we simply used the occurrence of spaces). However, one could easily improve this method by adding information—also possibly extracted automatically. For example, instead of a word-by-word alignment, a morpheme-by-morpheme alignment can be attempted, presupposing that we know about the morpheme separation of both languages. In the other direction, another possisble enhancement would be to mark frequent collocations in each language, and not align the individual words, but whole chunks of possible idiomatic expressions.

In contrast, instead of adding information beforehand, it is also possible to use the trans-co-occurrences (as, for example, extracted from Bible translations) for further linguistic analysis. For example, it turns out that (inflectional) morphological variations of the same root often occur together in the trans-co-occurrences (cf. *erretten/errette* and *Hand/Hände* in Table 2). This suggest that trans-co-occurrence statistics might also be used to investigate the inflectional structure of a language. However, all these suggestions are left for further research.

References

BIEMANN, CHIRISTIAN AND QUASTHOFF, UWE (forthcoming): Dictionary acquisition using parallel text and co-occurrence statistics, in: *Proceedings of NODALIDA-05, Joensuu, Finland.*

BIEMANN, CHRISTIAN; BORDAG, STEFAN; HEYER, GERHARD; QUASTHOFF, UWE & WOLFF, CHRISTIAN (2004a): Language-independent methods for compiling monolingual lexical data, in: *Proceedings of CicLING 2004, Seoul Korea.* [LNCS 2945]. Berlin: Springer, 215–228.

BIEMANN, CHRISTIAN.; BÖHM, K., HEYER, GERHARD, MELZ, R. (2004b): Automatically building concept structures and displaying concept trails for the use in brainstorming sessions and content management systems, in: *Proceedings of I2CS, Guadalajara, Mexico.* [LNCS 3473]. Berlin: Springer, ppp–ppp.

BROWN, PETER F.; COCKE, JOHN; DELLA PIETRA, STEPHEN A.; DELLA PIETRA, VINCENT J.; JELINEK, FREDRICK; LAFFERTY, JOHN D. ; MERCER, ROBERT L. & ROOSSIN, PAUL S. (1990): A statistical approach to machine translation, in: Computational Linguistics 16/2, 79–85.

BROWN, PETER E.; DELLA PIETRA, VINCENT J.; DELLA PIETRA, STEPHEN A. & MERCER, ROBERT L. (1993): The mathematics of statistical machine translation parameter estimation, in: *Computational Linguistics* 19/2, 263–311.

SAHLGREN, M. (2004): Automatic bilingual lexicon acquisition using random indexing of aligned bilingual data, in: *Proceedings of LREC-2004*, *Lisboa, Portugal*, 1289–1292.

VÉRONIS, JEAN (2000): From the Rosetta stone to the information society: A survey of parallel text processing, in: VÉRONIS, JEAN (ed.) *Parallel Text Processing: Alignment and Use of Translation Corpora.* [Text, Speech and Language Technology 13]. Kluwer: Dordrecht, 1–24.

Correspondence address

Michael Cysouw
Max Plank Institute for Evolutionary Anthropology
Deutscher Platz 6
D-04103 Leipzig
cysouw@eva.mpg.de