Dictionary Acquisition using Parallel Text and Co-occurrence Statistics

Chris Biemann and **Uwe Quasthoff**
NLP department
University of Leipzig
Augustusplatz 10-11
D-04109 Leipzig
{biem, quasthoff}@informatik.uni-leipzig.de

Abstract

We present a simple and efficient approach for deriving bilingual dictionaries from sentence-aligned parallel text by extending the notion of co-occurrences to a cross-lingual setting. Dictionaries are evaluated against gold standards and manually; the analysis accounts for frequency and corpus size effects.

1 Introduction

While more and more parallel text resources become available due to common regulations amongst different countries and the internationalization of markets, there is a need for translational lexicons that can keep pace with the daily creation of new terminology. The standard procedure amongst the Machine Translation Community to extract translation pairs from sentence-aligned parallel corpora can be outlined as follows (cf. (Melamed, 1996), (Moore, 2001) for an overview):

1. Define a measure which yields association scores between words of different languages
2. Calculate the association scores between all possible pairs
3. To obtain translations for a given word, sort its trans-lingual associations by score and take the top N or apply a threshold.

The crucial choice in the first step is to define a well-suited measure, the problem in step two is the effectiveness of the calculation, and in the last step the challenge is to cut the list at the right point.

In (Melamed, 1996), the problem of selecting inappropriate translations due to indirect associations is addressed: When e.g. *European* and *Council* often occur together in one language (here English) and *Europæiske* and *Råd* in the translation (here Danish), the measure yields a high association score between *Council* and *Europæiske* which is unwanted. He proposes a re-estimation method to alleviate this problem at the cost of long computation times. Other approaches like e.g. (Moore, 2001) or (Lee and Chang, 1993) rely on language-dependent resources like taggers, parsers or transliteration modules or use string similarity for preferring cognates.

We present an approach that is especially suited for large data sets, as it scales well with the corpus size - processing time is about 2 hours for a bilingual resource with about 28 million tokens in each language. Further, it does not make any assumptions on the languages and is therefore also suited for languages that are by no means close in terms of edit distance or character set (for languages without whitespace like Chinese and Japanese, a segmentation step is assumed in the preprocessing). We argue that a language independent approach like ours is a good baseline where more problem-specific methods (like using edit distances for cognates) can build upon.

Usually, co-occurrence statistics is used for large monolingual texts. It returns pairs of words that significantly often occur together within a predefined window. Implementations differ in the significance measure and the win-

dow size used. For our experiments, we use a log-likelihood-measure (which has been noticed to be adequate for the translation relation, see e.g (Tufiş, 2002)) and sentences as windows. This approach has proved to return pairs of semantically related words similar to human associations, cf. (Biemann et al., 2004).

For dictionary acquisition, we slightly modify the setting: Starting from a parallel corpus, we build a corpus containing bilingual *bi-sentences* built from pairs of aligned sentences. For technical reasons, we add language information to each word. Next, the co-occurrence statistics returns pairs of words which significantly often occur together in such bi-sentences. Here we consider only pairs of words having different language tags. Hence, for a given word we get an ordered list of translation candidates.

2 Methodology

Following the methodology of (Biemann et al., 2004), we adopt the machinery to compute co-occurrences of monolingual corpora (see *http://corpora.informatik.uni-leipzig.de* for co-occurrences in different languages) for the extraction of translation candidates: We mark the words in a bilingual sentence pair by source language and only regard co-occurrences between words of different languages. We call these significant trans-lingual associations *trans-co-occurrences*.
For two words A and B of different languages, each occurring a, b times in the corpus of bi-sentences and together in k of n bi-sentences in total, we compute the significance *sig(A,B)* of their trans-co-occurrence as follows:

$$sig(A,B) = \frac{\lambda - k\log\lambda + \log k!}{\log n} \text{ with } \lambda = \frac{ab}{n}.$$

To illustrate the necessary preprocessing, we give some examples for German-English bi-sentences as used in our experiments, words are marked by language:

- Die@de drogenfreie@de Gesellschaft@de wird@de es@de aber@de nie@de geben@de .@de But@en there@en never@en will@en be@en a@en drug-free@en society@en .@en
- Unsere@de Gesellschaft@de neigt@de leider@de dazu@de ,@de Gesetze@de zu@de umgehen@de .@de Unfortunately@en ,@en our@en society@en is@en inclined@en to@en skirt@en round@en the@en law@en .@en
- Zum@de Glck@de kommt@de das@de in@de einer@de demokratischen@de Gesellschaft@de selten@de vor@de .@de Fortunately@en ,@en in@en a@en democratic@en society@en this@en is@en rare@en .@en
- Ich@de sprach@de vom@de Paradoxon@de unserer@de Gesellschaft@de .@de I@en mentioned@en what@en is@en paradoxical@en in@en society@en .@en

In all pairs, *Gesellschaft@de* and *society@en* occur. Exactly this is used by the trans-co-occurrence calculation mechanism to produce the following top-ranked translation candidates:

- **Gesellschaft@de**: society@en (12082), social@en (342), our@en (274), in@en (237), societies@en (226), Society@en (187), women@en (183), as@en a@en whole@en (182), of@en our@en (168), open@en society@en (165)
- **society@en**: Gesellschaft@de (12082), unserer@de (466), einer@de (379), gesellschaftlichen@de (328), Wissensgesellschaft@de (312), Menschen@de (233), gesellschaftliche@de (219), Frauen@de (213), Zivilgesellschaft@de (179), Gesellschaften@de (173)

Note the large difference in significance between the first and all the other translation candidates. The significance values only have a local meaning: there is no global threshold for deciding whether candidates are good or bad, but within a ranked list of translation candidates for one word, they in fact tell the chaff from the wheat. To illustrate what

happens if words have several possible translations, the following example lists candidates whose significance values do not differ considerably:

- **kaum@de**: hardly@en (825), scarcely@en (470), little@en (362), barely@en (278), hardly@en any@en (254), very@en little@en (186), almost@en (88), difficult@en (68), unlikely@en (63), virtually@en (53), scarcely@en any@en (51), impossible@en (47), or@en no@en (40), there@en is@en (38), hardly@en ever@en (37), any@en (32), hardly@en anything@en (32), surprising@en (31), hardly@en a@en (29), hard@en (28), ...

- **hardly@en**: kaum@de (825), wohl@de kaum@de (138), schwerlich@de (64), nicht@de (51), verwunderlich@de (43), kann@de (37), wenig@de (37), wundern@de (25), man@de (21), drfte@de (17), gar@de nicht@de (17), auch@de nicht@de (16), gerade@de (16), berrascht@de (15), fast@de (14), berraschen@de (14), praktisch@de (13), ist@de (12), schlecht@de (12), verwundern@de (12), ...

Note that the methods produces a much longer list of candidates. Our machinery can deal with multi-words in a way that multi-words that are known beforehand can be treated as single units. For our experiments, we added multi-words that could be found in our evaluation dictionaries, e.g. "hardly any" in the examples above. To replace these manual resources, any approach for extracting collocations, e.g. (Smadia, 1993), could be applied in preprocessing; we did not undertake such efforts, which might be subject to further research.

3 Experiments

3.1 Data and resources

For evaluating our approach we use the multilingual, sentence-aligned Europarl Corpus (Koehn, 2002) available within the OPUS collection (*http://logos.uio.no/opus/*). The corpus consists of transcribed and translated parliament speeches of the EU parliament and covers a variety of topics. In each of the 11 languages, there are about 1 million sentences totalling about 28 million words. The corpus was neither tagged nor lemmatized.

Using machine-readable dictionaries for judging the quality of the obtained dictionary has a number of deficiencies. First of all, dictionaries are never complete, which leads to misjudging translations which might be correct, but not found in the dictionary. Second, a general-purpose dictionary might not suit domain-specific translations of the actual corpus. Third, if processing full forms, an automatic approach might give results in inflected form that cannot be found in a base-form dictionary, and lemmatizers or stemmers are not widely available for many of the European languages. Further, the differences between languages w.r.t compounding may lead to confusion in automatic evaluation. The results presented here can therefore only account for a lower bound of precision.
Comparing evaluations within different works is difficult. Many researchers lemmatize the corpus, some remove function words and some use word class information. What is considered correct or not for this task is also differing. In (Melamed, 1996), three kinds of 'correct' translations are given: pairs that are proper translations in *at least one context*, pairs that can be translations but have different parts of speech and pairs where one word is a possible part of the translation of another. Of course, this metric is much more relaxed then ours (which should not belittle his impressive results) and accepts entries you will never find in published dictionaries. Nevertheless, these entries are useful for machine translation or word alignment, see section 4 for an example.

3.2 Evaluation

We obtained machine-readable dictionaries from Freelang (*http://www.freelang.net*). Additionally we asked persons with very good knowledge of the respective language pair to judge the results without giving them

contexts. Due to availability of personal and dictionary resources, we tested our approach on the language pairs English-Danish, English-Dutch, English-Finnish, English-German, English-Italian, English-Portuguese and English-Swedish. Depending on the language, the method proposes at least three translation candidates for 34%-51% of total word types, which translates into 83%-89% coverage on tokens. For German-English, we used a larger dictionary in order to directly compare our approach to (Sahlgren, 2004), who uses random indexing instead of co-occurrence analysis but operates on the same data otherwise.

To alleviate the problem of full forms in the corpus as opposed to base forms in the dictionary to some extend, we use the following similarity measure: The similarity between two strings (words) V and W is defined as the number of similar prefix letters divided through the length of the longest string:

$$pfm(V,W) = \frac{\text{length of common prefix of } V \text{ and } W}{\max(length(v), length(W))}.$$

Although this measure is somewhat crude, it gives hints on how many correct, but inflected translations could be extracted. For evaluation, we checked the first three translation candidates for exact match (=1) and prefix match ≥ 0.6. This threshold produces almost no errors and takes flexations and sometimes part-of-speech change into account. It however does not capture Finnish case endings in all cases.

Table 3 in the appendix shows a randomly selected sample from the English-German data together with the maximum prefix match between the trans-co-occurrence candidates and the translations in the dictionary.

Table 1 contains precision values in % for all words that could be found in the dictionary for the first three translation candidates. Both translation directions are combined into one value.

For manual evaluation, we presented about 200 randomly selected words for each pair and let our language experts chose amongst correct, partially (part of a correct multiword, e.g. "Richter" instead of "Bundesrichter") and wrong. Table 2 depicts the results of manual evaluation.

language pair	1st candidate		2nd candidate		3rd candidate		1st, 2nd or 3rd cand.	
prefix match	=1	≥0.6	=1	≥0.6	=1	≥0.6	=1	≥0.6
en-sv	44.7	57.7	16.8	32.8	8.4	20.2	64.1	72.0
en-it	52.0	61.5	14.3	29.3	7.1	19.0	66.4	73.3
en-pt	56.6	66.7	12.8	27.3	6.4	17.4	70.6	78.7
en-nl	47.3	57.0	14.2	24.9	7.9	15.6	62.4	69.2
en-de	52.6	61.4	16.4	31.2	9.0	20.0	68.2	75.1
en-da	50.9	61.6	15.1	29.3	7.3	26.0	67.7	74.3
en-fi	34.9	51.3	17.3	34.26	10.4	25.9	52.0	65.9
total	46.1	58.0	15.8	30.9	8.6	21.2	62.2	71.3

Table 1: Dictionary precision for top-3 translation candidates in %

language pair	1st candidate			2nd candidate		
mode	corr.	part	both	corr.	part	both
de-en	64.5	19.6	84.1	25.5	22.4	47.9
en-de	52.8	26.4	79.2	35.9	25.5	61.4
da-en	55.5	21.8	77.3	12.7	18.2	30.9
en-da	56.8	27.0	83.8	27.9	25.2	53.1
sv-en	64.9	15.3	80.2	15.3	17.1	32.4
en-sv	67.6	13.5	80.1	36.0	10.8	46.8
nl-en	51.1	29.1	80.2	29.2	27.7	56.9
en-nl	56.6	18.9	75.5	32.0	24.5	56.5

Table 2: Precision in manual evaluation for first and second candidate in %

As expected, results of manual evaluation are higher than comparing to electronic dictionaries. Results demonstrate that our approach works in a language-independent fashion. Moreover, results for the first candidate are much higher than for the second, indicating that the association measure is appropriate.

For the remainder, we use German-English to determine the influence of word frequency and corpus size.

3.3 Influence of Frequency

As has been previously observed, precision of translation pairs is dependent on word frequency. (Sahlgren, 2004) uses co-occurrence information as well but reduces complexity by random indexing and stemming. He describes experiments for English-German with the same corpus and evaluates against the same large dictionary (*http://dict.tu-chemnitz.de/*). Intuitively, the more frequent a word is, the more reliable is the data for it, which should lead to higher results. The effect of using random indexing hence seems to prefer a certain frequency region, as Sahlgren obtained peak values in the absolute frequency range of 1000-5000.

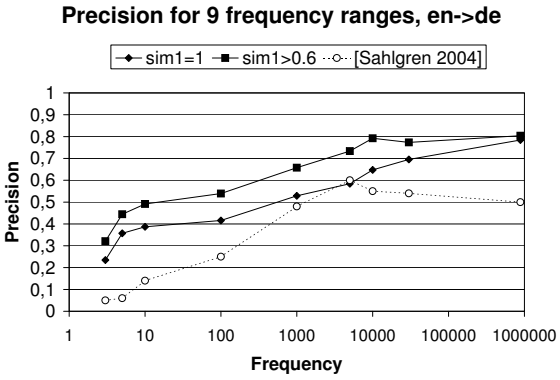

Figure 1: Dictionary precision for different frequency bands for English-German in comparison to (Sahlgren, 2004).

Our approach obtains a higher precision for all frequency bands, higher frequency leads to a higher precision, as figure 1 shows. Further, our method does not require brittle parameter tuning and does not introduce random.

3.4 Influence of Corpus Size

Another question we try to answer is the influence of corpus size. Of course, more parallel data means more evidence, so better results can be expected. But how little is necessary to start the process and what quality can we expect using much smaller parallel corpora?

We conducted experiments with English-German and gradually increased the number of parallel sentences from 500 to 1 million. The number of words, for which at least one trans-co-occurrence exists, grows a little less than linear in corpus size, as figure 2 indicates. This is the same progression as the total amount of word types.

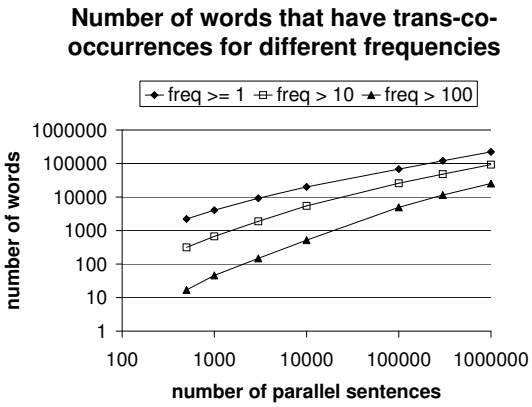

Figure 2: Number of words with trans-co-occurrences of different frequencies depending on corpus size

To measure whether corpus size influences the dictionary precision at absolute frequency count labels, we evaluated words of different frequency bands in our corpora of different sizes. We only took the highest-ranked trans-co-occurrence into account and accepted a translation as correct if a dictionary entry with prefix match ≥ 0.6 existed.The results as depicted in figure 3 indicate that precision is merely dependent on absolute frequency and not whole corpus size. The lower performance for the highest frequency band available in small corpora can be explained by the impossibility to give 1:1 equivalences of function words (e.g. (Catizone et al., 1989) give an example where German *auf* translates at least into English *for, in* or *on*). As these words constitute the topmost entries in a word list ordered descending by frequency, they deteriorate evaluation results if not many other words in the same frequency class are considered.

Another factor that might influence precision in dictionary acquisition is the (average) length of an aligned unit. Where we used the

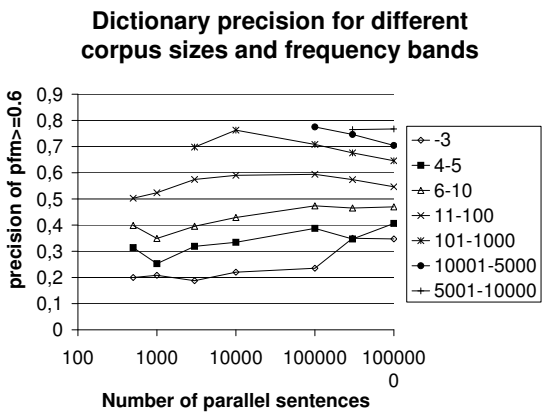

Figure 3: Dictionary precision for different corpus sizes and frequency bands containing at least 100 words

comparably long sentences of Europarl (about 28 words per unit), there exist versions of the Hansards corpus (as e.g. used by (Melamed, 1996)) that are aligned on a sub-sentence level in units of about 16 words.

4 Word Alignment

The ordered list of translation candidates per word can be used for word-to-word alignment of sentence-aligned corpora as follows: Given a sentence pair of L1 and L2, all words of L1 are linked to the word in L2 that can be found on the highest rank in the candidate list. In this way, even rare translations can be aligned, as the example in figure 4 shows: German *stellt* (usually English *puts*) and English *provides* (usually German *beliefert, beschafft*) are correctly linked, although *stellt* is ranked at position 15 in the candidate list for "provides". High frequency words as well as numbers are omitted in alignment.

In figure 5, another example demonstrating our approach's capability to do word alignment for long parallel units is depicted. It indicates that we can handle scrambling of word order naturally: In the first line in figure 5, the verb *setzen* int the final position of the clause is aligned to *setting* at third position. A simple heuristic could avoid multiple alignments of several determiners, as can be ob-

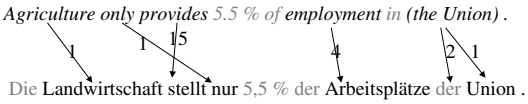

Figure 4: Grey words are not aligned, grey arrows indicate multiple alignments to the same word type. Numbers on arrows indicate the rank of the trans-co-occurrence. For a larger example, see figure 5.

served between *a* and *ein, eine, einen*. Finding several high-ranked trans-co-occurrences adjacent to each other (as with *Inkrafttreten* and *entry into force*) gives rise to the detection of multi-word units and their alignment. This breaks the 1:1 mapping assumption, which is especially not met when languages forming one-word compounds are involved. A more elaborate evaluation of this method on the word alignment of Bible texts can be found in (Cysouw et al., forthcoming).

5 Conclusion and future work

Co-occurrence statistics have proved to be useful to find translation pairs using parallel text, especially aligned sentences. Moreover, the need for large parallel texts is shown to extract large vocabularies. The next and more complicated question is to get rid of the sentence alignment and use only nearly parallel text. Here we have to replace the above bi-sentences by bi-texts an to calculate co-occurrences at bi-text level. While such texts are available at large scale (for instance, the multilingual Reuters corpus or multilingual web sites, see (Resnik and Smith, 2003)), processing is much more complex because it is quadratic in the length of the bilingual objects (i.e. texts instead of sentences). Further, longer units introduce more noise in the process.

6 Acknowledgements

The authors would like to thank Philipp Koehn for the preparation of the Europarl corpus and Wibke Kuhn, Ronny Melz and Pauliina Svensk for helping us with the manual evaluation.

References

Chris Biemann, Stefan Bordag, Gerhard Heyer, Uwe Quasthoff, and Christian Wolff. 2004. Language-independent methods for compiling monolingual lexical data. In *Proceedings of CicLING 2004, Seoul, Korea and Springer LNCS 2945 Springer Verlag Berlin Heidelberg*, pages 215–228.

R. Catizone, G. Russel, and S. Warwick. 1989. Deriving translation data from bilingual texts. In *Proceedings of the First International Lexical Acquisition Workshop, Detroit, USA*.

Michael Cysouw, Chris Biemann, and Matthias Ongyerth. forthcoming. Using strong's numbers in the bible to test an automatic alignment of parallel texts. *In: Michael Cysouw and Bernhard Wälchli (eds.) Parallel Texts: Using translational equivalents in linguistic typology. Special issue of Sprachtypologie und Universalienforschung.*

P. Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation.

Chun-Jen Lee and Jason S. Chang. 1993. Acquisition of english-chinese transliterated word pairs from parallel-aligned texts using a statistical machine transliteration model. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts, Data Driven Machine Translation and Beyond, Edmonton*, pages 96–103.

I. Melamed. 1996. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas, Montreal, Canada.*

Robert C. Moore. 2001. Towards a simple and accurate statistical approach to learning translation relationships among words. In *Workshop on Data-driven Machine Translation, 39th Annual Meeting and 10th Conference of the EACL, Toulouse, France*, pages 79–86.

P. Resnik and N.A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Magnus Sahlgren. 2004. Automatic bilingual lexicon acquisition using random indexing of aligned bilingual data. In *Proceedings of LREC-2004, Lisbon, Portugal.*

Frank Smadia. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1).

Dan Tufiş. 2002. A cheap and fast way to build useful translation lexicons. In *Proceedings of the 19th International Conference on Computational Linguistics COLING-02,Taipei, Taiwan.*

word	cand.1	pfm1	cand.2	pfm2	cand.3	pfm3
acute	**akuten**	0.66	**akute**	0.8	**akuter**	0.66
absolutely essential	*absolut*	0	*unbedingt*	0.166	**unbedingt notwendig**	0.1
essential	**wesentlichen**	0.83	**wesentliche**	0.909	ist	0
office	**Büro**	1	**Amt**	1	**Büros**	0.8
pollutants	**Schadstoffe**	1	**Schadstoffen**	0.916	Emission	0
expertise	**Fachwissen**	0	**Sachverstand**	1	**Sachkenntnis**	1
prescribed	**vorgeschrieben**	1	**vorgeschriebenen**	0.875	**vorgeschriebene**	0.93
means	**bedeutet**	1	**Mittel**	1	**heisst**	0.09
industrial goods	**Industriewaren**	0.64	*gewerbliche*	0	*Erzeugnisse*	0
bill	*Gesetzentwurf*	0.15	*Gesetzesentwurf*	0.133	**Rechnung**	1
approach	**Ansatz**	1	Konzept	0	**Vorgehensweise**	0
audit	*Prüefung*	0	**Audit**	1	**Rechnungsprüfung**	1

Table 3: Top three candidates for German-English sample with prefix match scores. Manually judged correct translations are marked in **bold**, part of translations *in italics*. Note the disagreement between automatic and manual evaluation.

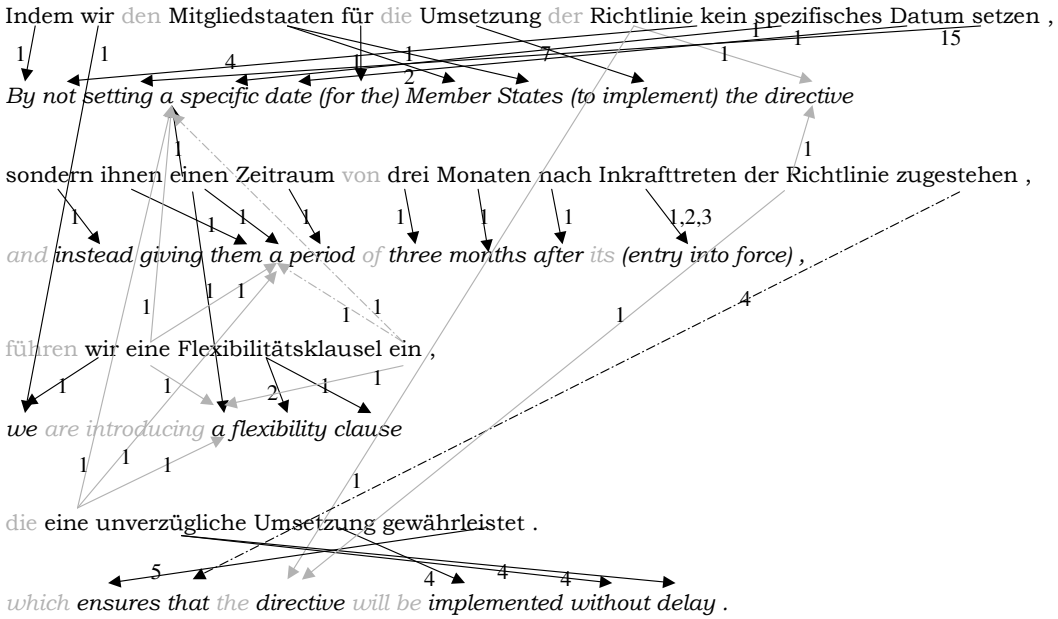

Figure 5: Second example for word alignment. Grey words are not aligned, grey arrows indicate multiple alignments to the same word type. Numbers on arrows indicate the rank of the trans-co-occurrence. The dashed arrow marks an alignment error. the only content word pair which is not aligned is the particle verb *führen .. ein* to its equivalent *introducing*.