ANCR – The Adjective-Noun Collocation Retriever (A Tool for Generating a Bilingual Dictionary from a German-English Parallel Corpus)

Günther Fliedl, Andreas Homa, Philippa Maurer-Stroh, Georg Weber
University of Klagenfurt
Universitätsstrasse 65-67
9020 KLAGENFURT
AUSTRIA
philippa.maurer-stroh@aon.at
Sebastian Maurer-Stroh
IMP - Institute of Molecular Pathology
Bioinformatics - Group Eisenhaber
Dr. Bohrgasse 7
1030 VIENNA
AUSTRIA

Abstract
ANCR is a tool which enables the user to automatically generate a bilingual dictionary of adjective-noun combinations from a sentence-aligned parallel corpus of German and English. The underlying idea of ANCR is to help translators and/or linguists retrieve correct translations of adjective-noun collocations without having to engage in time-consuming linguistic pre-processing tasks, which can depend on the availability of the appropriate software.

1 Introduction – Or the Motivation for ANCR

"Our knowledge of a language is not only a knowledge of individual words, but of their predictable combinations" (Stubbs 2001: 4). In fact, words in isolation can have different meanings from words combined as phrases. Take, *green light*, for example. The combination of the adjective *green* and the noun *light* does not only mean that there is a light which is green in colour, but also 'a permission for a project etc. to start or continue'. When it comes to contrastivity this phenomenon shows even more strongly. The English lemma POSITION can have various German translation equivalents, like *Lage*, *Standpunkt*, *Stellungnahme*, *Position*, etc. and it can only be correctly translated when its context, e.g. its preceding adjective, is taken into consideration. Therefore, *political position* is translated as *politische Position* and *geographical position* as *geographische Lage*, etc.

The way words combine in a language in order to produce natural-sounding, native-like speech or writing can be defined by the concept of collocation (Deuter et al. eds. 2002). Thus, *lasting peace* will be more acceptable in a native-speaker environment than *durable peace*, although both *lasting* and *durable* are translated as *dauerhaft* in German. Starting in the 1980s, high-storage capacities and increased data-processing speed have made it possible to capture this *language in use*. Sophisticated software enables users, e.g. lexicographers, to

view – and capture – words in their natural context. However, while nowadays "the use of text corpora is fairly firmly established in monolingual general-purpose, pedagogical and terminological lexicography, much remains to be done in bilingual lexicography." (Hartmann 1994: 292).

The only specialised corpus-based collocation dictionaries available are monolingual (cf. Deuter et al. eds. 2002) – there being very few bilingual collocation dictionaries of any kind available anywhere - and corpus-based general bilingual dictionaries still lack comprehensive phraseological treatment. Here is where the idea for ANCR was born.

ANCR is a tool which enables. The need for such a tool becomes particularly obvious when evaluating results of existing machine translation tools which are among the best in their area, as Figure 1 shows.

Source texts:

1. Die dauerhaften Entwicklungen der Weltwirtschaft führten letztendlich zu dauerhaftem Frieden.
2. Sustainable development ultimately led to lasting peace.

SYSTRAN (http://www.systranbox.com/systran/box)

Translation1: The durable developments of the world economy led finally to durable peace.
Translation2: Stützbare Entwicklung schließlich geführt zu dauerhaften Frieden.

PROMT ONLINE TRANSLATION (http://translation2.paralink.com/)

Translation1: At last the lasting developments of the world economy led to lasting peace.
Translation2: Aufrechtzuerhaltene Entwicklung schließlich geführt anhaltender Frieden.

LANGENSCHEIDT T1 5.0 (Demo-Version)

Translation1: Finally the durable <A[durable|permanent]> evolutions of the world economy led to lasting peace.
Translation2: Aufnehmbare Entwicklung führte letztlich zu haltbarem Frieden.

Figure 1: Translation results of existing machine translation tools

None of the tools tested provided the correct translation equivalents of both *dauerhafte Entwicklungen* and *dauerhaftem Frieden*, which are *sustainable developments* and *lasting peace.*

2 Machine Translation and Parallel Corpora

Translating from one language into another normally requires knowledge of two linguistic codes (cf. Koehn 2003). For a machine, such a task is impossible to perform without prior implementation of those two language systems – which can be a tedious and demanding undertaking.

Parallel corpora seem to be the solution. A parallel corpus is a collection of texts and their translations. The basic idea behind the use of parallel corpora in generating an adjective-noun collocation dictionary is that we can use the knowledge of thousands of translators which is stored in these translation corpora to get correct translation equivalents of usage-based, only on the level of convention, fixed adjective-noun collocations (cf. Brown 1997).

According to Koehn (2003), we can currently observe four major directions in machine translation: interlingua, transfer-based, example-based and statistical. While the former two are based on the implementation of linguistic knowledge, the latter two draw on parallel corpora as their reference source. Hence, both example-based and statistical tools let the machine extract translation equivalents from a bilingual corpus, avoiding time-consuming rule-writing and lexicon compilation processes. The difference between example-based and statistical machine translation lies in the fact that in statistical automatic translation probability calculations are added to the retrieval process (cf. Koehn 2003).

With ANCR we provide an example-based machine translation tool, the only implemented probability heuristic of which is (absolute) frequency since the concept of collocation can be, statistically, defined as the frequent co-occurrence of word-forms or lemmas (Stubbs 2001). This is where ANCR differs from recent approaches in the field of machine translation. A detailed description of machine translation can be found, inter alia, in Koehn (2003) and Al-Onaizan et al. (1999).

Albeit aiming at different outputs, we largely follow Brown's approach (1997) of an example-based extraction: ANCR is "knowledge free"; thus, no linguistic knowledge is required before the retrieval process. One of our main goals was to avoid linguistic pre-processing, such as tagging, lemmatising and (chunk) parsing – since these steps largely depend on available appropriate software for processing the language pair in question.

3 Methodology

Starting from a sentence-aligned parallel corpus of German and English (sentence-alignment tools are available free from the net), ANCR retrieves bilingual adjective-noun combinations and displays the results in a user-friendly, dictionary-like searchable interface.

We used the freely available *Europarl* corpus to carry out experiments. This corpus contains EU parliament proceedings in 11 languages and is already tokenised and sentence-aligned along Church and Gale's algorithm (Koehn 2002). The version of the German-English parallel sub-corpus we used contains 16,701,572 tokens on the German and 18,118,861 tokens on the English side, amounting to 734,327 sentence-pairs.

ANCR is written in perl. First, it loads both files, then converts the umlauts in the German corpus to separate vowel spelling so that the corpus can be processed on any machine. Next, ANCR reads in both files and extracts bi-grams from each line. Iterating the process line per line, the generated bi-grams of the source language, in our case German, are now juxtaposed with the bi-grams from the target language, the English, file. Once all sentence pairs have been processed, all identical bilingual bi-grams are added up since the hypothesis is that the more often two bi-grams co-occur in the same line in the parallel corpus the likelier they are to be translation equivalents. With the *Europarl* corpus we decided on a minimum joint frequency of 5, as this threshold provides the best results while not losing too much information (since there is no lemmatisation involved the threshold has to be set rather low because of the German language being highly inflected). Having generated a list of bilingual bi-grams with a minimum occurrence of 5, ANCR uses filters to post-process the results and then loads them into the Graphical User Interface (GUI).

One of the post-processing filters is a stopword list which is automatically generated from the parallel corpus. Tests have shown that, with a corpus the size of the *Europarl*, best

results are achieved when bi-grams containing the 100 most frequent words from each sub-corpus are excluded. The overall architecture of the tool can be seen in Figure 2.

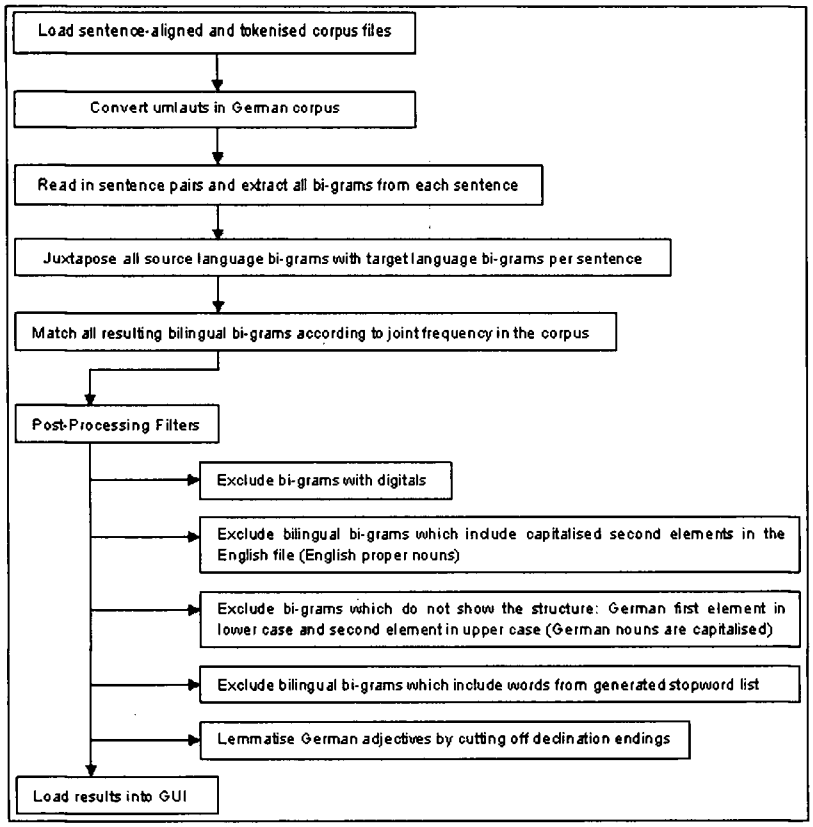

Figure 2: Architecture of ANCR

Despite the simplicity of the tool, it performs quite well. However, with a performance of 70% (validated by two native speakers – 720 correct translation bi-grams from 1029 retrieved) it is obvious that a number of enhancements have to be made.

4 Discussion, Future Work and Application

Starting from the 70% performance, we tested to what extend performance would be affected if we implemented a well thought out stopword list for both German and English as the error rate in retrieving German adjective-noun collocations and their English translation equivalents is mainly due to prepositions or pronouns preceding nouns in both files. This filter would enhance performance to 83.7% in the *Europarl* corpus.

As a next step we will test two methods to reduce the error rate resulting from the differences in German and English compounding. We created a perl script that correlates all

German nouns with a minimum length of eight characters with two adjacent words in the English file. In addition, we will combine that perl script with a list of English compounds (e.g. taken from WordNet). The evaluation of ANCR results run on *Europarl* shows that performance could be improved to 89.7% when compounds are correctly retrieved and allocated.

We also tested ANCR on Koehn's (2000) *de-news* corpus with 68,614 sentence pairs and on a manually compiled and sentence-aligned corpus of tourism texts with 344 sentence pairs. Both retrieval processes reveal approximately the same performance as the bilingual adjective-noun extraction from *Europarl*. Hence, ANCR can be applied to any parallel corpus of German and English.

Starting from the GUI, the parallel corpus files are loaded and ANCR performs its retrieval and filtering processes, displaying the results in such a way that the user can send queries starting from the adjective or noun in both languages, as can be seen in Figure 3.

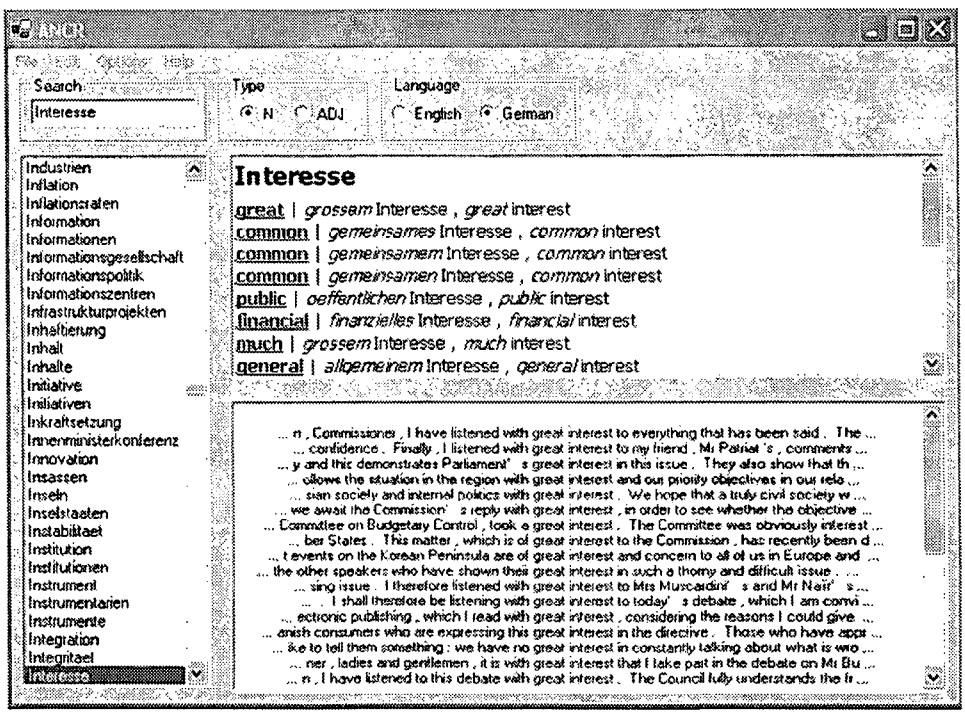

Figure 3: Example query with ANCR run on *Europarl*

In addition to the translations of adjective-noun collocations, ANCR also provides examples from the loaded parallel corpus to help the user correctly embed the combinations in texts.

5 Conclusion

With ANCR we created an example-based "knowledge free" tool which retrieves adjective-noun collocations from a sentence-aligned parallel corpus of German and English. The results are further processed to generate a bilingual dictionary of adjective-noun collocations which displays translation equivalents as well as example sentences from the corpus.

This tool can be used as a stand-alone programme to help lexicographers and translators in their decision-making process, but it can also be implemented in existing machine translation kits.

Acknowledgements

This project is kindly supported by the University of Klagenfurt.

References

Al-Onaizan, Y., Curin, J., Jahr, J., Knight, K., Lafferty, J., Melamed, D., Och, F. J., Purdy, D., Smith, N. A. and Yarowsky D. 1999. 'Statistical Machine Translation: Final Report' in *Johns Hopkins University 1999 Summer Workshop on Language Engineering, Center for Language and Speech Processing*, Baltimore, MD. Available at <http://www.isi.edu/~och/>.

Brown, R. D. 1997. 'Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation' in *Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation*, Santa Fe, NM. Available at <http://www-2.cs.cmu.edu/~ralf/papers.html>.

Deuter, M., Greenan, J., Noble, J. and Phillips, J. (eds.) 2002. *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Hartmann, R. R. K. 1994. 'The Use of Parallel Text Corpora in the Generation of Translation Equivalents for Bilingual Lexicography', in W. Martin et al. (eds.) *Euralex 1994 Proceedings*. [Amsterdam:] N.p.: 291-297.

Koehn, P. 2000. 'German-English Parallel Corpus "de-news", Daily News 1996-2000'. Available at <http://www.isi.edu/~koehn/de-news>.

Koehn, P. 2002. 'Europarl: A Multilingual Corpus for Evaluation of Machine Translation'. Unpublished, available at <http://www.isi.edu/~koehn/europarl>.

Koehn, P. 2003. *Noun Phrase Translation*. Unpublished Ph.D. thesis, University of Southern California. Available at <http://www.isi.edu/~koehn/publications/>.

Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford and Malden, MA: Blackwell Publishers.