Phrase processing for detecting collocations with KoKS

Norman Kummer
Joachim Wagner
University of Osnabrück
Institute of Cognitive Science
D-49078 Osnabrück

norman@VauDePe.de
jowagner@uos.de

1. Introduction

KoKS stands for [Ko]rpusbasierte [K]ollokations-[S]uche, in English “corpus based search for collocations”. The aim of the KoKS project was to build a database which contains bilingual (for the first step: German-English) pairs of phrases, classified according to whether they are collocations or not. The bilingual phrases are put to use in an application presented in section 3.8.

The system was developed within a student project (s. Erpenbeck et al. 2002) at the University of Osnabrück. It processes texts and their translations to create hypotheses of phrase correspondences employing an initial bilingual lexicon. Once detected, the "new" phrases will extend the bilingual lexicon if they pass a further statistical filter. This way the system's database grows steadily.

In section 2, we will explain KoKS's definition of collocation. The formulation was chosen to fit our bilingual approach. The next section focuses on the components of KoKS. Then, in section 4, we briefly present results and discuss them. Section 5 outlines a few applications that are related to computer assisted language learning (CALL) and machine translation (MT). Next, in section 6, we try to relate our approach to other works published so far. Finally, section 7 briefly discusses open problems.

2. Which phrases are considered to be collocations within KoKS?

KoKS's considers a phrase to be a collocation if it cannot be translated word for word. If a compositional translation of a German phrase cannot be found in its English counterpart sentence, it is likely that a collocation in accordance to Breidt's collocation definition (Breidt 1995) has been found:

> “[...] collocations shall refer only to word combinations with a lexically (rather than syntactically or semantically) restricted combinatory potential, where at least one component has a special meaning that it cannot have in a free syntagmatic construction”

If a phrase is not translated word for word on a regular basis, we assume that there must be a component that has a special meaning, i.e. that there is a collocation. In other words, a compositional translation is interpreted as an indicator of compositional semantics. Consider the example “k*ick the bucket”*. Its meaning can not be derived compositionally and the German translation *“ins Gras beißen”* is not word for word.

Two types of errors which are of varying importance to different applications must be distinguished. On the one hand, there are several collocations satisfying Breidt's definition that can be translated word for word from English to German or vice versa. In this case, KoKS would be unable to detect them. On the other hand, phrases may be freely translated even if a literal translation is also possible. In such a case, KoKS will erroneously detect a collocation.

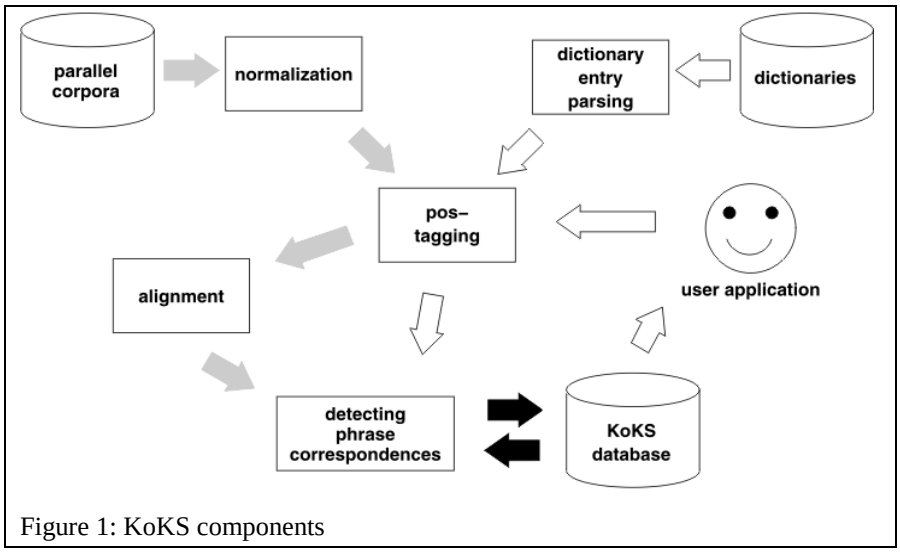

Figure 1: KoKS components

3. The KoKS system

The aim of the KoKS project was to build a system that detects collocation candidates by looking for a missing literal translation of a phrase in the corresponding sentence and employing simple statistical tests. To obtain suitable data, a long chain of acquisition and pre-processing is necessary: corpora and dictionaries have to be normalized and tagged, and paragraphs, sentences and phrases have to be aligned. This results in pairs of phrases which are stored and linked in a database. These are considered to be collocation candidates. This set of candidates still includes phrases that are translated literally, so, according to our definition, they are not collocations.

The system does not identify collocations. But it shows a collocativity measure. Thus, the KoKS system cannot really distinguish between these two groups. KoKS orders the phrases in a continuum.

In this section we will describe only the main components of the system shown in figure 1 taken from Koch (2001). Details can be found in Erpenbeck et al. (2002). The components are presented according to the order in which the information flows through the system.

3.1 Used Corpora and Dictionaries

Our approach relies on the analysis of parallel corpora. However, freely available bitexts are a limited resource. The following corpora (table 1 outlines the most important facts of these corpora) were obtained and processed:

1. DE-News[1]: This corpus consists of news reports broadcasted over a period of several years, the original language was German. They were translated into English within a voluntary project by non-professional human translators. The format is ASCII/HTML; the majority of the texts are short.

2. EU-publications[2]: This corpus contains press releases, news, political documents and contracts. The format is HTML-like; the texts are also short.

We also tried to make use of four further corpora: Bible, Linux HowTo, NATO publications and the Verbmobil corpus. However, several problems occurred during

[1] http://www.isi.edu/~koehn/publications/de-news/
[2] http://europa.eu.int/rapid/start/welcome.htm

processing e.g. difficult aligment and non-standard formats. Because the EU-publications seemed to be sufficient, we decided not to go further in solving those problems.

Corpus	Files	Size KB	Lines	Words	Characters
De-News	2,214	14,542	274,959	1,912,206	13,454,497
EU	23,610	93,683	1,580,780	11,513,213	83,867,545
Total	25,824	108,225	1,855,739	13,425,419	97,322,042

Table 1: Basics statistics of our corpora

In addition to corpora the system employs dictionaries[3]. Table 2 shows the number of original entries and of those that are calculated after the DEP process described in the next subsection.

Dictionary	Number of entries	Number of entries after DEP-process	Direction
Ding	124423	151684	Ger->Eng
Tyler Chamber	9749	10105	Eng->Ger
Unknown	31856	36180	Eng->Ger
LQL (Byrd 1989)	45825 80534	184940	Ger->Eng Eng->Ger

Table 2 Statistics of the dictionaries

3.2 Normalization / DEP

The corpora we use are in different text formats, e.g. HTML, SGML and PDF. Normalization means that they are converted into the same text format. Only sentence and paragraph boundaries are kept if available.

The huge number of files (over 26,000 in January 2002) is managed with XML files that describe the corpus files in a consistent manner. In this way the appropriate normalization module for each file can be chosen automatically.

Much more work has to be done to process dictionaries. The formats vary and do not conform to any standard. Within the KoKS project, tools were developed that parse dictionary entries (dictionary entry parsing, DEP) and produce a two column output. Table 3 shows a few lines of DEP in-/output.

The further processing that is described in the following subsections is required for both normalized corpora and pre-processed lexical entries. In the latter case, alignment is skipped, because the rows of the tables already represent the correct alignment of the two languages.

DEP input (example) German::English

Pöbel::mob,populace,rabbie,riffraff
Bank,Damm,Ufer,Böschung,Reihe::bank

Normalized DEP output

German	English
Pöbel	mob
Pöbel	populace
Pöbel	rabble
Pöbel	riffraff
Bank	bank
Damm	bank
Ufer	bank
Böschung	bank
Reihe	bank
gedrängt wie die Sardinen	packed like sardines

Table 3: DEP in-/output

3.3 POS tagging & lemmatization

The IMS Tree-Tagger (Schmid 1994) is applied to perform part-of-speech tagging and lemmatisation, it uses the Stuttgart-Tübingen tagset STTS[4] for German and the Penn Treebank Project[5] tagset for English.

3.4 Alignment of sentences

Sentence alignment is a very important aspect of the KoKS project. It is the basis of

[3] Ding: http://www.tu-chemnitz.de/dict
Tyler and Chamber: http://www.june29.com/IDP/
LQL: http://www.cl-ki.uni-osnabrueck.de/~ulf/uni/ws98-99/lexikon/

[4] ftp://www.ims.uni-stuttgart.de/pup/corpora/stts_guide.ps.gz

[5] ftp://ftp.cis.upenn.edu/pup/treebank/doc/cl93.ps.gz

a good phrase alignment. We decided to develop our own aligner, because most of the software modules could be useful both in the sentence alignment and phrase alignment (see below).

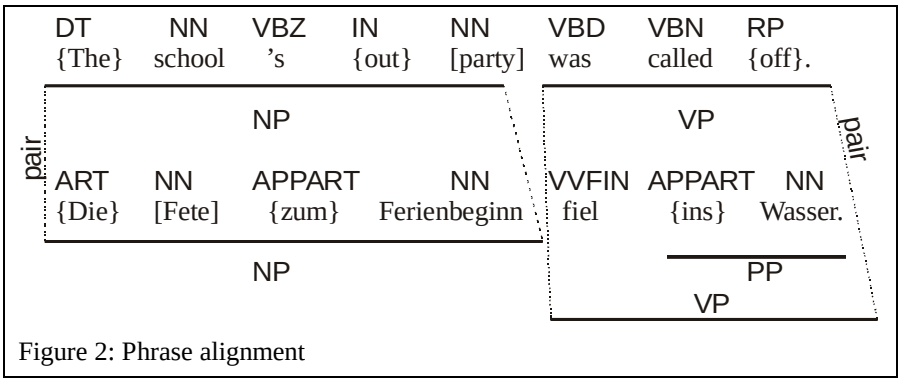

Figure 2: Phrase alignment

The sentence aligner uses lexical knowledge to calculate a distance measure of sentences. Our measure combines three methods. Firstly, it consults the dictionary looking up lemmas that have been annotated by the IMS Tagger. The translation correspondences that are found are counted. Secondly, the distance measure searches for trigram correspondences within the remaining open-classed words. Both steps just consider words that belong to an open class, i.e. their POSs is noun, verb, adjective or adverb. The distance values are calculated by dividing the number of counted correspondences by the maximal number of open classed words. Thirdly, our measure compares the number of close class words. Likewise, sentence length in characters are compared using the method presented in (Gale and Church 1993). The overall distance value is the weighted sum of the three parts. For further details and examples the reader should consider (Erpenbeck et. al 2002).

A sentence alignment is a sequence of beads that unites corresponding sentences. An alignment bead can contain more than one sentence in each language. This is necessary if a translator split a sentence or joined some sentences.

An optimal alignment is calculated using the A* - graph-searching algorithm. It is used to find the cheapest path in the distance matrix of all sentence pairs. This path represents the optimal alignment.

3.5 How does KoKS align the phrases?

Phrase hypotheses are generated based on the POS tags which the IMS-Tagger annotates. KoKS matches all connected subsequences of tags with a table of predefined tag sequences ordered by syntactic category.

This table is arranged in different ways for English and German. In order to get German tag sequences, we employed a monolingual corpus that was chunk-parsed using IMS[6] tools. We used the chunks to identify sequences of POS tags. These sequences can be queried directly with the IMS tool CQP[7]. The English tag sequences are extracted from our own POS-tagged KoKS corpus. Each sequence has to conform to one of the following rules (notated as a regular expression[8]):

1. NP := [DT] N ([IN] N)*
2. PP := {IN|TO} NP
3. VC := (V [TO])*
4. VP := VC [NP] [PP]

The names of these rules stand for the corresponding syntactic category in which the sequence is stored.

[6] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

[7] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/

[8] [X] the term X occurs 0 or 1 time, { } lists alternatives terms, X* = the term X may occur 0, 1 or more times, ( ) = concatenation of terms

Now let us go back to the description of the phrase alignment process. Figure 2 adapted from Koch (2001) gives an example of how phrase alignment works in KoKS:

1. Mark words that have irrelevant tags ({…})
2. Mark words that have translations in the other sentence ([…])
3. Construct tag sequences by category (——)
4. Pair all tag sequences with equal categories (.....)
5. Pair leftover words

We just use words that belong to an open class as starting points of looking for matching tag sequences. This is sufficient, because all four types of tag sequences defined above do contain an open class POS.

If for each open class words of a tag sequence its translation according to the KoKS dictionary can be found in the other sentence, the phrase will not be paired with any other phrase that satisfies the same condition. Of course, such phrases should not be paired with any other phrase. This is a task for improvements of the system.

The system aligns all English and German phrase hypotheses that belong to the same category. Words that do not belong to any tag sequence are also paired in order to take the chance to find additional correspondences.

While looking for a counterpart in the corresponding sentence, KoKS skips phrases and word pairs which are already stored in the database. Any bilingual pair of phrases found this way is stored in the database either as a new entry or as one more instance of a phrase pair.

3.6 Detecting Collocations with KoKS

For each phrase pair we count how many times it has been found within the phrase alignment process. Furthermore, we calculate a measure of collocativity. Here we benefit from our definition of collocations (see section 2). Whether a phrase can be translated word for word or not can be measured with the distance measure described in section 3.4. This statistical information is used to obtain relevant phrase pairs. A phrase pair is considered relevant if it occurs in least a adjustable number of sentence pairs. The results are ordered by collocaticity.

3.7 The database

The database consists of tables which are highly inter-connected. It provides information on which sentences contain specific tokens, and it counts how many times a phrase was found in the corpus. (A phrase has to be found at least a certain number of times – the value can be changed – to be considered interesting for KoKS.) The system can reconstruct the original sentence from which a word or phrase was taken. The sentence alignment is represented with the help of shared key numbers that identify alignment beads. Phrases are stored like sentences.

All of this is the system's knowledge about sentence and phrase alignment. The KoKS database can be queried by SQL. This option provides a good and powerful basis for specialized and KoKS-system-independent lookups. So the database can be used easily in other or future developments.

3.8 Demo-Application

The KoKS project also developed a web-application in the CALL-context, that helps a L2-learner to understand phrases, which cannot be translated literally (collocation-like multi-word phrases).

If the learner requests sentence clarification, KoKS will query its database for all phrase hypotheses that can be

produced as described in section 3.5. The learner can than choose a phrase hypothesis and read the associated translations. Example sentences are also available to aid understanding.

Up to now, KoKS does not reliably state whether a phrase is a collocation or not. Thus, the learner is confronted with all relevant phrases the system has stored. But they are ranked by our collocativity measure. For example, if the German phrase "ins Wasser fallen" has been identified within the input sentence, the system shows the phrase multiple times ordered by descending collocativity, because different translations are stored in the KoKS database. The learner can choose one of these alternatives. In the example, "[0.210] ins Wasser gefallen – fell into the water" refers to the sentence

> *Das Kind **ist ins Wasser gefalllen**. - The child fell into the water.*

whereas the phrase "[1.000] ins Wasser gefallen – was cancelled" refers to

> *Die Party **ist ins Wasser gefallen**. - The party was cancelled.*

The fact that the KoKS system also presents the first alternative in which the phrase is not a collocations, might be considered to be a lack. But we think that is not the case for this CALL-embedded application, because the learner gets the requested help. In the example, the system presents both sentence pairs, so the learner could come aware of the fact that there are two different use cases of the phrase *"ins Wasser gefallen"*, the first sense is literal and the second is collocative.

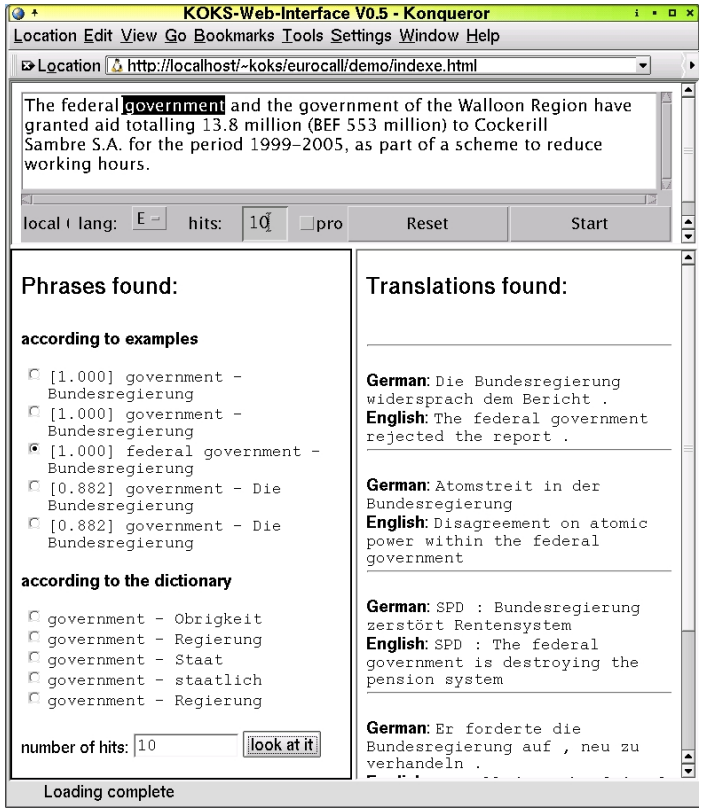

Figure 3: The KoKS Demo-Application

4 Results

Currently our phrase aligner has processed all sentence of our corpus up to a length of 20 tokens. These are approximately 48000 sentence pairs. We did start analysing short sentences to reduce the number of phrase combinations and run time. Furthermore, we hoped to improve the quality of phrase hypotheses.

4.1 Phrase Alignment

It very important that the phrase pairs are correctly aligned. The meaning of the phrases must be similar if not identical. Figure 4 shows the precision of the phrase alignment for different minimal numbers of references.

4.2 Collocativity Measure

A lot of phrase pairs with high collocativity are not collocation in Breidt's sense. Never-

#examples	German	English	collocativity	alignment
65	Bundesregierung	federal government	1.000	good
40	soll	is expected	1.000	bad
36	soll	is supposed	1.000	good
28	Kohl	Chancellor Helmut	1.000	good
27	Bundesrat	Upper House	1.000	good
26	teilte	was announced	1.000	good
24	Bundesregierung	German government	1.000	good
23	Herzog	German President	1.000	good
...				
19	Landgericht	Regional Court	1.000	good
...				
7	Fischer	Federal Foreign Minister	0.967	good
7	Aussichten	Extended forecast	0.962	good
10	Schröder	Federal Chancellor	0.960	good
7	Mitgliedstaat	Member State	0.958	good
16	Fischer	Foreign Minister	0.955	good
15	Gesamtkosten	total cost	0.952	good
...				
14	Landgericht	The Regional	0.882	near
13	keine Einwände	Commission	0.882	bad
13	Bundeskabinett	federal cabinet	0.882	good
13	einem Zeitungsinterview	newspaper	0.882	near
12	Beihilfeintensität	The aid	0.882	near
12	Mitgliedstaaten	The Member	0.882	near
...				
13	einem Zeitungsinterview	a newspaper interview	0.536	good
9	Das Bundesverfassungsgericht	The Federal Constitutional Court	0.533	good
15	die Europäische Union	European Union	0.529	good
8	EU-Komission	EU Commission	0.529	good
11	Die Grünen	The Green	0.519	good
33	Bundesregierung	Federal Government	0.500	good
...				
14	Die Europäische Gemeinschaft	The European	0.286	good
12	der Europäischen Union	the European	0.286	good
8	Das Europäische Parlament	The European	0.286	good
8	Die Bundesregierung	The Federal government	0.286	good
7	Der Bundesgerichtshof	The Federal High	0.286	good
...				
8	Die Union	The Union	0.000	good
8	die CDU	the CDU	0.000	good
8	Die FDP	The FDP	0.000	good
7	der CDU	the CDU	0.000	good
7	Der Bundesgerichtshof	The Federal	0.000	near

Table 4: Detected phrase pairs ordered by collocativity and number of references

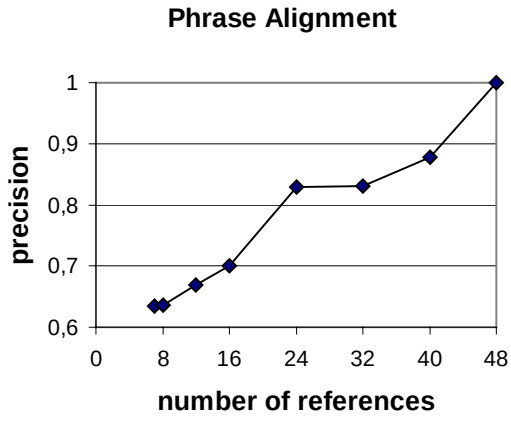

Figure 4: Phrase alignment precision

theless, most of them are important translations that are missing in our current dictionaries (see table 4). In a few spot checks, we were unable find an example even in commercial dictionaries.

5. What can KoKS be used for?

5.1 Translation of collocations (intelligent lookup)

Nobody can deny the fact that electronic dictionaries can be improved. In a text-understanding scenario, they often fail to take account of the context of a word. Collocations especially are not recognized. This leads to problems to translate occurrences of collocations. For example, consider the German sentence

> *"Der Vortrag fiel wegen einer Erkältung des Dozenten ins Wasser."*

Here an adjunct interrupts the collocation *"fiel ins Wasser"* which complicates its detection.

A master thesis currently being written that aims to teach the KoKS system, to detect occurrences of collocations like *"fiel ins Wasser"* in a user's text input, even if a word or phrase stands between the parts of the collocation. To aid understanding, examples of occurrences of the collocation in original corpora will also be presented. These are the steps involved in processing the user's input (see figure 5 adapted from Koch, 2001):

1. Tag the input sentence
2. Mark words that have irrelevant tags ({…})
3. Construct tag sequence (——) paying special attention to the marked word
4. Search for translations in the database, especially for the canonical form of the collocation, i.e. *"ins Wasser fallen"*, which originate from dictionaries
5. Collect example sentences from database

5.2 Providing examples

KoKS can provide examples of occurrences of phrases in their original environment. No time-consuming search is necessary, because the KoKS database stores references to the sentences in which phrases were detected.

Furthermore, the system can perform full text searches that are known from traditional search engines. This is done with the help of an index of sorted tri- and bigrams. Of course, both searches are based on the lemmas annotated by the IMS-tagger and not on the tokens themselves.

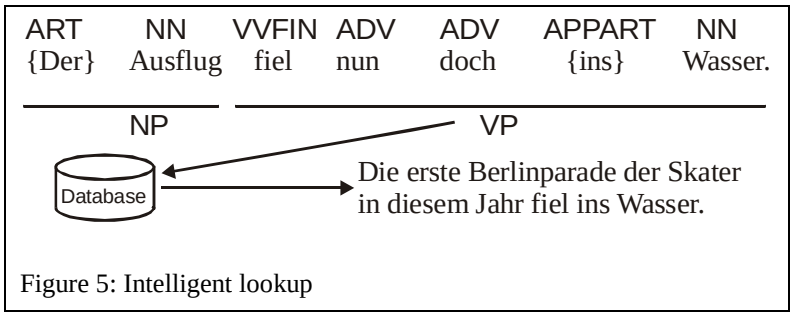

Figure 5: Intelligent lookup

5.3 Translation memory (TM)

TM systems assist a human translator in choosing a translation that is consistent with previously-made translation decisions. They can also save time if the text is repetitive (Benis 1999). Traditional TM systems do not explicitly store phrases. They can only actively help a human translator if the whole sentence to be translated is similar to a stored one. In these systems the translator has to request a full text search to find a phrase.

The KoKS database stores all phrase correspondences detected by the system. While the user translates a new sentence, the system could search for known phrases on its own. The measure of collocativity could make the translator aware of special uses of components. The KoKS system is prepared to incrementally process bilingual data. So it should be easy to import additional material as soon as the translator has finished a paragraph.

6 Related work

A good overview on corpus pre-processing, alignment and detection of collocation gives (Somers 2001). He refers to several approaches to identify phrase pairs, for example to Daille who uses tag sequences as we do.

(Wu 1995) introduces invers transduction grammars (ITG) to align the phrases of bilingual sentences. ITG parse trees imposes a shared structure on both sentences. The ITG formalism allows to constrain the possible phrase matchings.

An approach to discover phrases that are not literally translated is presented in (Melamed 97). He employs translation models that make few assumption about the languages in the bitexts. In this context, another interesting work on lexicon extraction is (Tiedemann 2000). (Orasan 2000) splits sentences into clauses using machine learning techniques.

7. Outlook and Open problems

Further experiments in intelligently combining statistics and collocativity measure are neccessary in order to achieve a better separation of collocations and phrases that are translated word for word.

It depends in particular on the lexicon as to whether a collocation can be detected. For example, “*starker Raucher - heavy smoker”* will not be identified as a collocation, if “*heavy”* is listed in the translations of “*stark”* (strong).

Our idea to generate phrase hypotheses has to be improved. The tables of tag sequences must be enlarged and verified. Tag sequences coverage may be improved by inducing them across our aligned corpus in a way similar to the ideas in (Yarowsky 2001).

The concept has to be adapted to be able to account discontinous phrases. We want to achieve this without syntactic parsing. One idea might be to formulate constraints that inserted words must satisfy.

References:

Benis, Michael (1999): Translation Memory. In: Bulletin of the Institute of Translation and Interpreting, issue April - May 1999

Breidt, Elisabeth (1995): Extracting V-N-Collocations from Text Corpora: A Feasibility Study for German. In: Proeedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, Ohio.

Byrd, Roy "LQL user notes: An informal guide to the lexical query language." Technical report, IBM T. J. Watson Research Center, New York, 1989.

Erpenbeck, Arno; Koch, Britta; Kummer, Norman; Reuter, Philipp; Tschorn, Patrick & Wagner, Joachim (2002): KoKS - Korpusbasierte Kollokationssuche. Abschlussbericht. Universität Osnabrück. http://www.cl-ki.uni-osnabrueck.de/~koks/main/bericht/

Gale, William A. & Church, Kenneth W. (1993): "A programm for aligning sentences in bilingual corpora". Computational Linguistics 19: S. 75–102.

Klavans, Judith L. & Tzoukermann, Evelyne. Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons. Machine Translation, 1996.

Koch, Britta (2001): KOKS4CALL. Presentation slides. http://www.cl-ki.uni-osnabrueck.de/~koks/main/praesentation/3/

Melamed, I. Dan (1997): Automatic Discovery of Non-Compositional Compounds in Parallel Data. Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 97-108

Orasan, Constantin (2000): A hybrid method for clause splitting in unrestricted English texts, In: Proceedings of ACIDCA'2000, Monastir, Tunisia

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: International Conference on New Methods in Language Processing

Somers, Harold (2001): Bilingual Parallel Corpora and Language Engineering. Anglo-Indian workshop "Language Engineering for South-Asian languages" (LESAL), (Mumbai, April 2001).

Tiedemann, Jörg (2000): Extracting Phrasal Terms using Bitext. Proceedings of the Workshop on Terminology Resources and Computation, held in conjunction with LREC 2000, Athens/Greece

Wu, Dekai (1995): Grammarless extraction of phrasal translation examples from parallel texts. In Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, Leuven, Belgium.

Yarowsky, David; Ngai, Grace & Wicentowski, Richard (2001): Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. http://citeseer.nj.nec.com/yarowsky00inducing.html