

The use of multi-level Annotation and Alignment for the Translator

Mihaela Vela

Applied Linguistics,
Translation and Interpreting
Saarland University,
Germany

m.vela@mx.uni-saarland.de

Silvia Hansen-Schirra

Computational Linguistics &
Applied Linguistics,
Translation and Interpreting
Saarland University,
Germany

hansen@coli.uni-sb.de

Abstract

In translation practice, typological differences between languages can pose problems in such a way that the translator has to compensate a source language structure which does not exist in the target language. It is, however, not always easy to find an adequate translation equivalent for such a lacking structure. The aim of this paper is to show how a multiply annotated and aligned corpus can be used as a translation memory for such typologically driven problems, exploiting the linguistic enrichment of the corpus. It is discussed how an existing translation corpus with high-quality translation and alignment is converted into a database and how this database can be exploited as a large on-line resource displaying various translation options for lexico-grammatical problems.

1. Introduction

Up to now the annotation of translation corpora, i.e. their linguistic enrichment, has been carried out in order to empirically investigate the properties of translated text. On the other hand, practical translators also work with large amounts of translated texts, the enrichment of these parallel texts, however, being mostly limited to sentence alignment. The use of these aligned texts in translation memories is again limited to string-based queries (see section 2).

There are, however, translation problems which are due to typological differences between languages. For translation training and practice it would, consequently, be good to have a database where many examples of these typological characteristics and their translations into other languages can be found. For English and German, raising structures, extractions and deletions are among others problematic constructions (cf. Hawkins 1986). Because here, the search space for a translational choice is rather wide, finding the German translation equivalent for such a construction is therefore a notorious problem for which a parallel concordance can provide help. While with a raw text corpus we can only formulate string searches, an annotated and aligned corpus allows us to query for lexico-grammatical patterns (see section 4). Additionally, we show how our corpus alignment can help to create multilingual term bases for translation practice and training. The advantage of this technique is that the translation candidates are extracted from published translations, i.e. language in use, and are thus more comprehensive and inventive than dictionary entries are.

The research described here is part of a pilot project called KOALA¹ for which we use the CroCo corpus. The design of the corpus as well as its multidimensional annotation and alignment are described in section 3. It is shown how the multiply annotated and aligned corpus is imported into a database and how this database can be exploited to solve typical translation problems for English and German.

Another crucial issue when dealing with large amounts of source language texts and their translations is the preservation of the meta-information of the texts (i.e. information on the author, translator, nationality and mother tongue of author and translator, publication date, etc.). In order to store and manage this kind of information for each text in the corpus, we developed a graphical user interface, CroCo-Meta. This tool allows the annotation of important meta-information in a user-friendly and fast manner (see section 5).

Finally, we conclude the paper with summing up the advantages of our methodology and discussing some directions for future research (see section 6).

2. Corpora in translation studies and translation practice

2.1 Corpora in translation studies

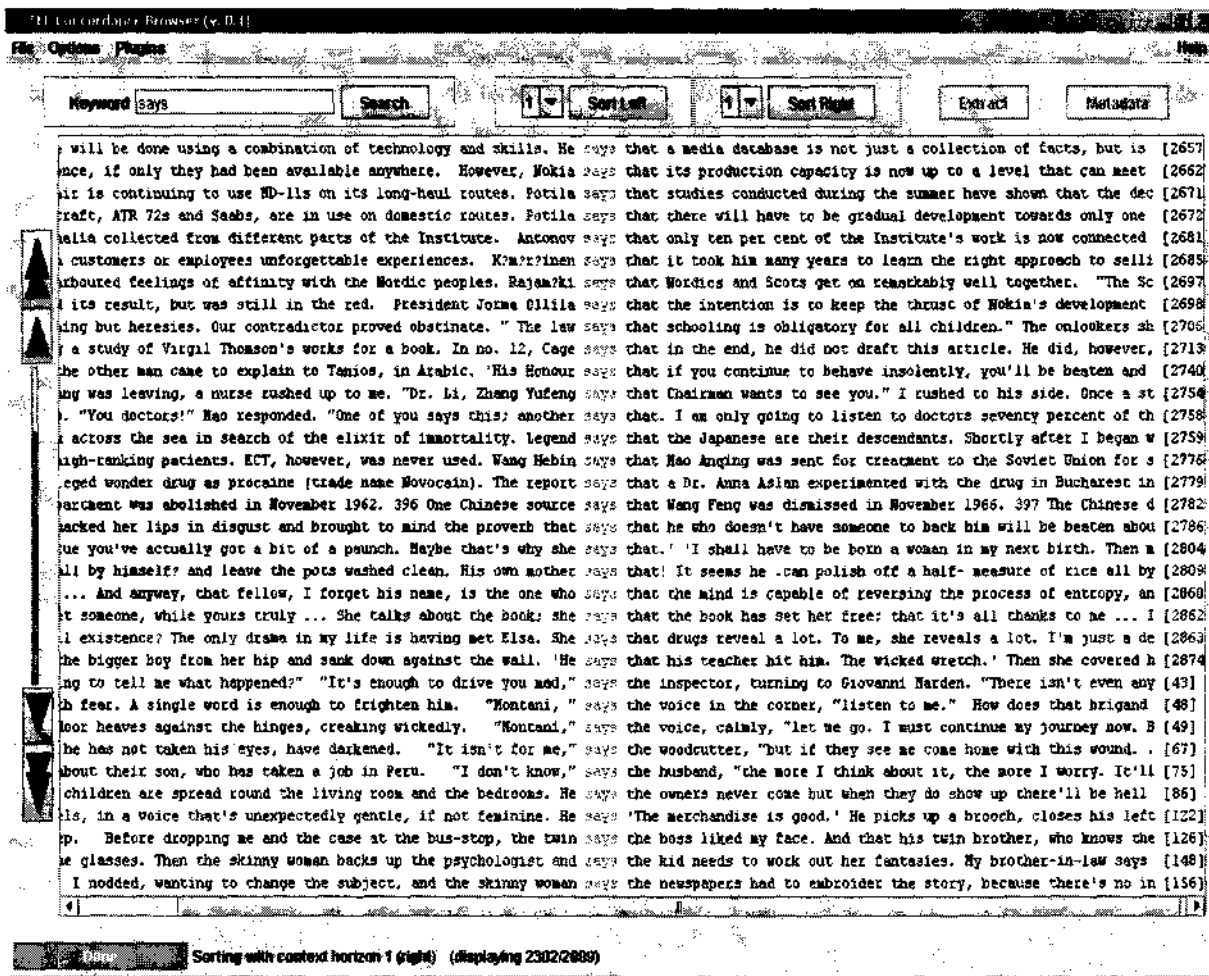
In translation studies, parallel corpora are used to analyse characteristic patterns of translations. Within this context Baker (1996) formulates the following hypotheses on the universal features of translations: explicitation (translations are more explicit than originals), simplification (translations are easier to understand and more readable), normalisation (translations strongly adhere to the usage norms of the target language) and levelling out (translations are more alike than the individual texts in a corpus of originals). These hypotheses are tested in several studies using TEC and BNC as comparable corpora: Laviosa-Braithwaite (1996) tests simplification and levelling out by analysing average sentence length, lexical density and type-token ratio. Baker/Olohan (2000) find evidence of explicitation in TEC investigating the use of *that*-connectives in contrast to zero-connectives of the reporting verbs *say* and *tell* (see Figure 1).

Furthermore, Olohan tests the hypothesis of explicitation on the basis of the use of contractions (Olohan 2003) and optional syntactic elements in translations (Olohan 2004). Hansen (2003) finds evidence of normalisation in a tagged version of TEC and explains this result with the help of a psycholinguistic test. Based on comparable corpora, translation universals, such as explicitation, simplification or normalisation, are also tested for many other languages (e.g. Bernardini/Zanettin 2004 or Kujamäki/Mauranen 2004).

Using a corpus of English source language texts, German translations, and comparable German originals, Hansen/Teich (1999) and Teich/Hansen (2001b) investigate the above mentioned translation features explicitation, simplification, normalisation and levelling out. The use of a combined parallel-comparable corpus allows them to identify the influence of the source language texts on the translations and on the target language. Teich (2003), for instance, finds shining-through (the typical language use of the source language “shines through” in the German translations), thus detecting a tendency contrary to normalisation.

Parallel corpora are used for the investigation of information structure in English and German texts (e.g. Doherty 1999), thematic structure (Hasselgard 1998), information packaging (e.g. Steiner 2002, 2004 for English-German and Fabricius-Hansen 1999 for German-English and German-Norwegian) and again for the investigation of explicitation in English and Norwegian parallel texts (Johansson 1995).

¹ <http://fr46.uni-saarland.de/koala/>



Finally, corpus-based methods are used in order to investigate translator's style (Baker 2000, Olohan 2004), creativity in translation (Kenny 1998) as well as intercultural issues (Mauranen 1997). Such analyses could positively influence translators' training and practice as well as translation criticism.

2.2 Corpora in translation practice

When speaking about multilingual corpora as reference resources for professional translators, a useful resource, which is freely available and easily accessible for many different languages, is the World Wide Web (cf. Kilgarriff 2001). Lexical look-up can, for example, be initiated via a search engine like Google. The matches are displayed in the form of links to web sites, which can be investigated. An easier way to linguistically explore the Web is the tool WebCorp², which displays the matches in the form of a concordance, a KWIC output. Another useful application of the Web as a translation aid is to search for multilingual web sites, i.e. translated web sites, from which parallel texts can easily be downloaded, aligned and used as a parallel corpus or a translation memory.

Using parallel corpora in translation practice, we have to distinguish between the translation of highly repetitive texts (such as manuals or instructions) and the translation of creative writing (e.g. fiction). In the context of translation of software manuals, for instance, it can be worthwhile to compile previously translated manuals, align them, and store them in a database, which can then be used for reference. Such translation memories are usually equipped with alignment and terminology management tools (e.g. Translator's Workbench by Trados, Déjà vu by Atril or Star Transit) and are thus useful both for terminology look-up and for pre-translation of phrases and even for whole sentences.

² <http://www.webcorp.org.uk>

But even for texts that are not highly repetitive, multilingual and especially parallel corpora can support the translation process. A parallel corpus can be employed as a multilingual lexical resource, being more comprehensive and diverse than dictionaries. A multilingual comparable corpus can be used for exploring register use as well as typological differences. This is extremely helpful for the translation of special purpose texts and the acquisition of highly specialised terminology since term banks and glossaries can be built up easily (Bowker/Pearson 2002). If a corpus is linguistically annotated, it can also be used to help solving grammatical or semantic translation problems. For the translation of literary texts, the investigation of a comparable corpus can reveal the personal style of an author or translator that may be incorporated in the translation.

The basic idea of using corpora in translator training is that a parallel corpus consists of a more comprehensive and diverse variety of source language items and possible translation solutions than a dictionary could ever display (cf. Zanettin/Bernardini/Stewart 2003). Thus, in translator training, parallel corpora are explored for terminology look-up on the one hand (Pearson 2000, Danielsson/Ridings 2000, Maia 2003 as well as Bowker/Pearson 2002 in a Languages for Specific Purposes (LSP) context) and for teaching the usage of collocations (Teubert 2003 and Barlow 2000 using parallel corpora) as well as register- and typology-specific patterns of the target language on the other (Pearson 2003 and Bowker 1999 using comparable corpora). Hansen/Teich (2002) show how an English-German translation reference corpus annotated with part-of-speech tags can be used to look up not only lexical items but also grammatical structures. Furthermore, a parallel corpus can show translation students and language learners how to deal with translation problems (see Pearson 2003 for English-German and Johansson/Hofland 2000 for English-Norwegian) and how to avoid typical mistakes (see, for instance, Vintar/Hansen 2005 analysing cognates in parallel texts).

One interesting approach to providing students with insights into possible translation strategies is to collect several translations of one and the same source language text (cf. Teubert 2001). A similar scenario is introduced by Johansson (2003), where translations into several languages are collected from one and the same source language text. Here, students are able to learn to which degree the linguistic structures of the source language text can be preserved in the target language or how they have to be transferred according to the norms of the target language.

Another common method of teaching and studying translation is the use of learner corpora. Here, several translations of one and the same source language text produced by students are collected. A very easy way to collect such learner texts is to submit and store them electronically as proposed by Bowker/Bennison (2003) in their work on the Student Translation Archive. Also, possible translation errors or peculiarities could be tagged and explored in such a way that the students can learn from the translation behaviour of other learners and translators (cf. Malmkjær 2003).

3. The CroCo Corpus as linguistically enriched translation memory

For the experimental project KAOLA we used the CroCo Corpus (cf. Hansen-Schirra et al. 2006), a linguistically annotated and aligned corpus for German and English. The CroCo corpus was collected for the investigation of the translation property of explicitation for the language pair English - German and consists of English originals, their German translations as well as German originals and their English translations. Both translation directions are represented in eight registers. Biber's calculations, i.e. 10 texts per register with a length of at least 1,000 words, serve as an orientation for the size of the sub-corpora (cf. Biber 1993). Altogether the CroCo Corpus comprises one million words. Additionally, reference corpora are included for German and English. The reference corpora (see Figure 2) are register-neutral including 2,000 word samples from 17 registers (see Neumann/Hansen-Schirra 2005 for more details on the CroCo corpus design).

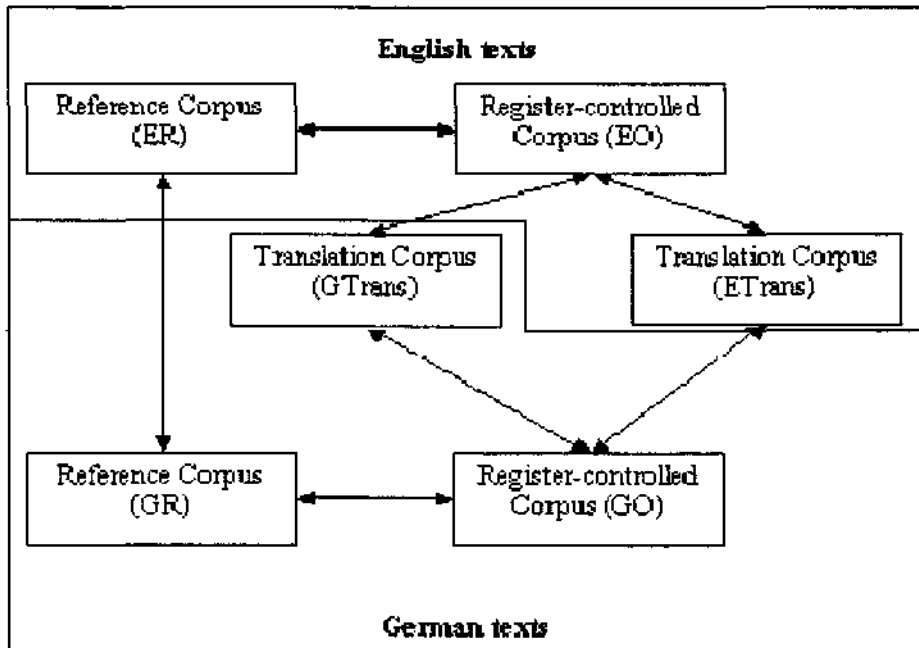


Figure 2: Corpus design in CroCo

The CroCo Corpus is tokenised and annotated for part-of-speech, morphology, phrasal categories and grammatical functions. Furthermore, the following (annotation) units are aligned: tokens, clauses and sentences. The annotation and alignment steps are described in section 3.1. The transformation of the annotation and alignment into a MySQL database is described in section 3.2.

3.1 Annotation and Alignment for KOALA

In this section we describe the type of information annotated and aligned in the CroCo corpus, since this information is the basis for the experiments run in the KOALA project.

For each text in the corpus the annotation covers different levels. Thus, each kind of annotation (part-of-speech, morphology, phrase structure, grammatical functions) is realized in a separate layer. An additional layer is included which contains comprehensive meta-information in separate header files for each text in the corpus (see section 4).

At each annotation level and for each text there is a base file consisting of the indexed units in the text. Index and annotation layers are kept separate using XML stand-off mark-up based on XCES³. The base file at each level contains the segmentation of the text at the specific level. At token level the index file consists of the indexed words, at chunk, clause and sentence level of the indexed chunks, clauses and sentences. In turn, the index files at chunk, clause and sentence level refer to the index files at token level.

The first layer to be presented here is the tokenisation layer. Tokenisation is performed for both German and English by TnT (Brants 2000), a statistical part-of-speech tagger. As shown in Figure 3 each token annotated with the attribute **strg** has also an **id** attribute, which indicates the position of the word in the text. This **id** represents the anchor for all XPointers pointing to the tokenisation file by an **id** starting with a “t”. The file is identified by the **name** attribute. The **xml:lang** attribute indicates the language of the file, **docType** provides information on whether the present text is an original or a translation.

³ <http://www.xml-ces.org>

```

<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE document SYSTEM "token.dtd">
<document xmlns:xlink="http://www.w3.org/1999/xlink"
  name="EO_SHARE_001.tok.xml" xml:lang="en" docType="ori">
<header xlink:href="EO_SHARE_001.header" />
  <tokens>
    ...
    <token id="t4" strg="Fiscal" />
    <token id="t5" strg="2002" />
    <token id="t6" strg="was" />
    <token id="t7" strg="a" />
    <token id="t8" strg="very" />
    <token id="t9" strg="challenging" />
    <token id="t10" strg="year" />
    <token id="t11" strg="for" />
    <token id="t12" strg="the" />
    <token id="t13" strg="entire" />
    <token id="t14" strg="industry" />
    ...
  </tokens>
</document>

```

Figure 3: Tokenisation and indexing

The second layer annotated is the part-of-speech layer, which is provided again by TnT⁴. The token annotation of the part-of-speech layer starts with the **xml:base** attribute, which indicates the index file it refers to. The part-of-speech information for each token is annotated in the **pos** attribute, as shown in Figure 4. The attribute **strg** in the token index file and **pos** in the tag annotation are linked by an **xlink** attribute pointing to the **id** attribute in the index file.

```

<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE document SYSTEM "tagEnglish.dtd">
<document xmlns:xlink="http://www.w3.org/1999/xlink"
  name="EO_SHARE_001.tag.xml">
  <tokens xml:base="EO_SHARE_001.tok.xml">
    ...
    <token pos="jj" xlink:href="#t4" />
    <token pos="mcl" xlink:href="#t5" />
    <token pos="vbdz" xlink:href="#t6" />
    <token pos="at1" xlink:href="#t7" />
    <token pos="jb" xlink:href="#t8" />
    <token pos="vvg" xlink:href="#t9" />
    <token pos="nnt1" xlink:href="#t10" />
    <token pos="if" xlink:href="#t11" />
    <token pos="at" xlink:href="#t12" />
    <token pos="jb" xlink:href="#t13" />
    <token pos="nnj1" xlink:href="#t14" />
    ...
  </tokens>
</document>

```

Figure 4: PoS tagging

Morphological information is particularly relevant for German due to the fact that this language carries syntactic information within morphemes rather than in separate function words like English. Morphology is annotated in CroCo with MPro, a rule-based morphology tool (cf. Maas 1998). This tool works on both languages. The encoding of the morphological information is analogue to the part-of-speech encoding shown in Figure 4.

Moving up from the token unit to the chunk unit, the same pattern as for the tokens is repeated. The chunks are also indexed each chunk having and an **id** attribute as shown in Figure 5.

⁴ For German we use the STTS tag set (Schiller et al. 1999), and for English the Susanne tag set (Sampson 1995).

```

<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE document SYSTEM "chunk.dtd">
<document xmlns:xlink="http://www.w3.org/1999/xlink"
name="EO_SHARE_001.chunk.xml">
<chunks xml:base="EO_SHARE_001.tok.xml">
<chunk id="ch1">
  <tok xlink:href="#t1"/>
  <tok xlink:href="#t2"/>
  <tok xlink:href="#t3"/>
</chunk>
<chunk id="ch2">
  <tok xlink:href="#t4"/>
  <tok xlink:href="#t5"/>
</chunk>
<chunk id="ch3">
  <tok xlink:href="#t6"/>
</chunk>
<chunk id="ch4">
  <tok xlink:href="#t7"/>
  <tok xlink:href="#t8"/>
  <tok xlink:href="#t9"/>
</chunk>
...
</chunks>
</document>

```

Figure 5: Chunk indexing

The phrase structure annotation (see Figure 6) assigns the **type** attribute to each phrase chunk identified by MPro. XPointers link the phrase structure annotation to the chunk index file. It should be noted that in CroCo the phrase structure analysis is limited to higher chunk nodes.

```

<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE document SYSTEM "ps.dtd">
<document xmlns:xlink="http://www.w3.org/1999/xlink"
name="EO_SHARE_001.ps.xml">
<chunks xml:base="EO_SHARE_001.chunk.xml">
<chunk type="none" xlink:href="#ch1"/>
<chunk type="np" xlink:href="#ch2"/>
<chunk type="vp_fin" xlink:href="#ch3"/>
<chunk type="np" xlink:href="#ch4"/>
...
<chunk type="pp" xlink:href="#ch8"/>
<chunk type="clause" xlink:href="#ch9"/>
<chunk type="np" xlink:href="#ch10"
...
</chunks>
</document>

```

Figure 6: Phrase structure annotation

The annotation of grammatical functions is again kept in a separate file and is analogue to the phrase structure annotation.

On clause and sentence level there is only information about the segmentation of the text in clauses and sentences. As shown in Figure 7, each clause consists of a list of XLinks to tokens in the index file denoted by the **xml:base** attribute. The same approach applies also for the sentences.

```

<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE document SYSTEM "clause.dtd">
<document xmlns:xlink="http://www.w3.org/1999/xlink"
name="EO_SHARE_001.clause.xml">
<clauses xml:base="EO_SHARE_001.tok.xml">
<clause id="c11">
  <tok xlink:href="#t4"/>
  <tok xlink:href="#t5"/>
  <tok xlink:href="#t6"/>
  <tok xlink:href="#t7"/>

```

```

<tok xlink:href="#t8"/>
<tok xlink:href="#t9"/>
<tok xlink:href="#t10"/>
<tok xlink:href="#t11"/>
<tok xlink:href="#t12"/>
<tok xlink:href="#t13"/>
<tok xlink:href="#t14"/>
<tok xlink:href="#t15"/>
<tok xlink:href="#t16"/>
<tok xlink:href="#t17"/>
<tok xlink:href="#t18"/>
<tok xlink:href="#t19"/>
<tok xlink:href="#t20"/>
</clause>
<clause id="c12">
<tok xlink:href="#t21"/>
<tok xlink:href="#t22"/>
<tok xlink:href="#t23"/>
<tok xlink:href="#t24"/>
<tok xlink:href="#t25"/>
</clause>
...
<clause id="c117">
<tok xlink:href="#t168"/>
<tok xlink:href="#t169"/>
<tok xlink:href="#t170"/>
<tok xlink:href="#t171"/>
<tok xlink:href="#t172"/>
</clause>
</clauses>
</document>

```

Figure 7: Clause segmentation

In the examples shown so far, the different annotation layers linked to each other belonged to the same language. By aligning words, clauses and sentences, the connection between original and translated text is made visible. For the purpose of the CroCo project word alignment is realised with ATLAS (cf. Schrader 2006) an alignment tool which combines linguistic and statistical approaches. Clauses are aligned manually with the help of MMAX II (cf. Müller/Strube 2003), a tool allowing assignment of own categories and linking units. Finally, sentences are aligned using Win-Align, an alignment tool within the Translator's Workbench by Trados (cf. Heyn 1996).

The alignment procedure produces three new layers: token alignment, clause alignment and sentence alignment and follows like the annotation layers the XCES standard. Figure 8 shows how clause alignment is encoded. The **trans.loc** attribute locates the clause index file for the aligned texts. Furthermore, the respective language as well as the n attribute organising the order of the aligned texts are given. We thus have an alignment tag for each language in each clause pointing to the clause index file.

```

<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE document SYSTEM "clauseAlign.dtd">
<document xmlns:xlink="http://www.w3.org/1999/xlink"
name="E2G_SHARE_001.clauseAlign.xml">
<translations
xml:base="CROCO_CORPUS/ENGLISH2GERMAN/GTrans/SHARE/ANNOTATED/clause/">
<translation trans.loc="EO_SHARE_001.clause.xml" xml:lang="en" n="1"/>
<translation trans.loc="GTrans_SHARE_001.clause.xml" xml:lang="ge" n="2"/>
</translations>
<clauses>
<clause>
<align xlink:href="#c11"/>
<align xlink:href="#c11"/>
</clause>
<clause>
<align xlink:href="#c12"/>
<align xlink:href="#undefined"/>

```



```

</clause>
<clause>
  <align xlink:href="#c13"/>
  <align xlink:href="#c13"/>
</clause>
<clause>
  <align xlink:href="#c14"/>
  <align xlink:href="#c14"/>
</clause>
...
</clauses>
</document>

```

Figure 8: Alignment of clauses

The alignment of tokens and sentences in the corpus follow the same principle. Additionally, phrase alignment can be derived from word alignment and syntactic functions can be mapped automatically across the parallel corpus.

3.2 The KAOLA Database

Since we want to show how linguistic information can facilitate the query for typological differences of languages, the annotation and alignment work described in the previous section is the basis for such experiments. In order to have a fast search we decided to convert the annotation and alignment presented above in tables of a MySQL database.

id	string	pos	lemma	alignedWith
1	Dear	ii	dear	1
2	Shareholder	nn2	shareholder	3
3	1999	nc	1999	7
4	has	vbz	have	8
5	proved	vvn	prove	8
6	a	ai	a	9
7	difficult	ii	difficult	10
8	yet	tr	yet	12
9	successful	ii	successful	13
10	year	nn1	year	14
11	for	ii	for	15
12	out	appg	out	16
13	corporation	nn1	company	17
14		yi		18
15	Difficult	ii	difficult	19
16		yh		20
17	because	cs	because	21
18	we	ppst2	we	22
19	had	vhd	have	30
20	to	to	to	29
21	make	vv0	make	29
22	some	dd	some	24
23	more	dai	more	23
24	fundamental	i	fundamental	25
25	changes	nn2	change	26
26	in	i	in	0
27	the	ai	the	0
28	group	nn1	group	0

Figure 9: Database tables for English tokens

All available information on token level, such as tokenisation, part-of-speech and lemma including word alignment is written into the tables in the database (see Figure 9). The English tokens in Figure 9 are indexed, each index having assigned a string, a lemma, a part-of-speech tag and an index for its German equivalent. At chunk level the tables are filled with information about chunk type and the

grammatical function it fulfils. Similar to the XML encoding in the corpus, the MySQL tables for chunks are strongly connected to the information at token level. Analogously, the clause and sentence segmentation as well as their alignment is transformed into tables connected to the token tables in the MySQL database.

This type of storage allows us for a more easy and fast method to query the corpus. Additionally, a query interface with a menu-like, predefined set of queries can be connected to the database, allowing also for non-experts to query the corpus.

4. Multi-level annotation and alignment for solving typical translation problems

In many cases, typological differences between languages can be translated straightforward without any problems. Different word order or grammatical morphologies are not considered as major translation problems. There are, however, typological differences that are problematic for the translation process. Typically, these are constructions which exist in one language but which do not exist or are rarely used in the other. This means for the translation of such constructions that the translator has to compensate them in the target language. It is, however, not always easy to find an adequate translation equivalent. For this reason, a database with translation examples of typological differences can help to solve translation problems.

In the following, we explain the advantage of such a resource (see section 3) on the basis of Hawkins' (1986) descriptions for English and German. He states that English is far more productive concerning cleft sentences, raising constructions and deletions. Therefore, in the process of translating these constructions into English, compensations have to be found. Cleft constructions are, for instance, a typical feature of the English grammatical system. While they do exist in German as well, German has other options of realizing information distribution patterns, e.g., by word order variation. In our annotated and aligned corpus database, cleft constructions can be found by querying the following pattern:

```
word="it" FOLLOWED BY lemma="be" (FOLLOWED BY syntactic
function="complement" (INCLUDING pos="relative pronoun"))
```

Applying this query to our database, we find the following translation pairs⁵:

- (1) It is this ownership that we truly believe helped our employees to drive toward success, despite the challenges of this year. -- Mit dieser Beteiligung am Unternehmen im Rücken haben unsere Mitarbeiter nach unserer Überzeugung maßgeblich zum Erfolg des Unternehmens trotz der großen Herausforderungen dieses Jahres beigetragen.
- (2) It is to everyone's credit that we accomplished so much - the best year ever in our combined history. -- Dem Einsatz aller ist es zu verdanken, dass wir so viel erreicht haben.
- (3) In fact, it was their persistence through some very challenging days in 1998 that helped us end the year with such strong momentum. -- Tatsächlich ist es ihrem Durchhaltevermögen während einiger sehr kritischer Tage 1998 zu verdanken, dass wir das Jahr dann doch noch mit einem solch gewaltigen Erfolg beenden konnten.

Here, two options of translating English clefts into German are shown: The first example is nominalised in the German version, whereas in example 2 and 3 German infinitival constructions are chosen. In the latter examples a lexical pattern for translating clefts becomes visible: the translators used "es ist jemandem/etwas zu verdanken, dass" (it is somebody/something to thank that") for the translation of both English cleft sentences. This might be an indicator for a good translation strategy

⁵ The examples discussed in this paper are taken from our English-German sub-corpus of business communication.

for clefts. To find other strategies, it can be specified in the database query whether the clefted element is translated and thus aligned with a German adverbial, a German subject or other realisations.

Another interesting construction for which English is more productive than German is raising. Raising constructions can be found by querying the following pattern:

```
syntactic function="finite verb" (FOLLOWED BY syntactic
function="direct object" (REALISED THROUGH phrasal category="clause"))
```

With this query subject-to-subject raising can be retrieved as can be seen in the following examples:

- (4) We continue to benefit from the strong natural gas market in North America. -- Wir profitieren weiterhin von einem starken Erdgasmarkt in Nordamerika.
(5) We defined the minivan, and will continue to do so. -- Wir haben den Minivan erfunden und wir werden auch künftig neue Marktsegmente definieren.
(6) ... and attracting the best talent possible as we continue to grow our business. --- ... und werben zur Erweiterung unseres Geschäftes die besten Talente an, die wir nur finden können.

Here, one possible translation strategies become obvious: the meaning of the verb "continue" which occurs very frequently in the sub-corpus of business communication is translated by using temporal adverbials in German ("weiterhin" and "künftig" in examples 4 and 5). Additionally, in example 6 the verbal group is transformed into a nominal structure which seems to be a typical translation strategy for English-German.

According to Hawkins, substitutions and deletions occur more frequently in English than in German. We search for deletions using for example the following patterns:

```
phrasal category="prepositional phrase / noun phrase"
NOT INCLUDING pos="noun"
```

This query displays all nominal phrases in which the nominal head is deleted.

```
phrasal category="sentence" INCLUDING 2 * syntactic function="FIN"
AND 1* syntactic function="SUBJ"
```

This query displays all sentences with two finite verbs where one of the subjects is deleted.

```
Pos="conjunction" NOT ALIGNED ON WORD LEVEL
```

Here, we look for conjunctions which are optional in one language and therefore not translated. Surprisingly, we found more deletions in the German translations than in the English originals:

- (7) After the interviews, I told our employees that I wanted Baker Hughes to improve from being a good company to become a great one. -- Nach den Gesprächen sagte ich den Mitarbeitern, dass ich Baker Hughes von einer guten Firma zu einer erstklassigen machen wolle.
(8) We want to thank shareholders for your confidence, and we will continue to do everything possible to reward that confidence. -- Wir möchten den Aktionären für das uns entgegengebrachte Vertrauen danken und werden weiterhin alles Erdenkliche tun, dieses Vertrauen zu belohnen.
(9) Today, integrated functional departments, and shared ideas and technologies, are significantly improving everything we make, the way we do business, and the way we serve our customers - as this report shows. -- Heute verbessern integrierte Bereiche und der Austausch von Ideen sowie Technologien nicht nur unsere Produkte, sondern auch die Art, wie wir unsere Geschäfte führen und unseren Kunden dienen.

Example 7 shows a German prepositional phrase where the nominal head is deleted. In the English original a substitution is used to express the same meaning. Since substitution does not work for

German, the deletion is one strategy to translate this structure. In example 8, we find two German verbs (“danken” and “tun”), the second subject is, however, deleted. In the English original the subject “we” is repeated for the second verb. The same phenomenon can be observed in example 9: The English original repeats the words “the way we”, which is deleted in the German translation. In both examples, German is more elliptical expressing the cohesive links implicitly, whereas English uses repetitions expressing the lexical cohesion more explicitly.

Another application of our word alignment is that it can be used to create a bilingual dictionary or a bilingual term base. We can extract, for examples, specific categories (like verbs, nouns) from our aligned corpus. However, also other combinations are possible: German tends, for instance, to use many compounds, which use to be multiword units in English. Such multiword-noun alignment can consequently be extract from the corpus⁶:

```
<item>
  <lemma>silk handkerchief</lemma>
  <category>multiword</category>
  <language>English</language>
  <translations>
    <translation>
      <lemma>einstecktuch</lemma>
      <category>noun</category>
      <language>German</language>
      <confidence>0.2</confidence>
    </translation>
  </translations>
</item>
```

Figure 10: English-German word alignment

Another interesting issue is the transformation of word classes during the translation process. In order to detect these examples verb-noun alignments can be extracted:

```
<item>
  <lemma>sneaking</lemma>
  <category>verb</category>
  <language>English</language>
  <translations>
    <translation>
      <lemma>schleichtour</lemma>
      <category>noun</category>
      <language>German</language>
      <confidence>0.2</confidence>
    </translation>
  </translations>
</item>
```

Figure 11: English-German word alignment

On the basis of such a bilingual term database, the register as well as language conventions for using verbal or nominal constructions can be investigated very easily.

5. Documentation of meta-information using CroCo-Meta

This section is an excursion to the annotation and storage of meta-information in a corpus. This is an important issue because only a good and transparent documentation of meta-information enables accurate and efficient queries.

CroCo-Meta is a GUI developed in order to provide a user-friendly storage of meta-information for a text in a corpus. The meta-information that can be written in the CroCo-Meta form as well as the

⁶ The following to examples are taken from our sub-corpus fiction.

format in which the meta-information is saved are based on the TEI guidelines (Sperberg-McQueen/Burnard 1994).

CroCo-Meta allows the annotator to process the meta-information element by element. He can either insert the required information into the corresponding field or select the appropriate entry from a predefined list. The tool then creates automatically the header file and saves the information from the form fields in the CroCo-header format as shown in Figure 15.

The CroCo-Meta Form is divided into four parts: *File Description*, *Encoding Description*, *Annotation Description* and *Alignment Description*. The *File Description* part of the GUI depicted in Figure 12 encodes information about the file itself, the sub-corpus to which it belongs, the translation, the author, publication, the title, as well as a short register analysis⁷. The register analysis contains information about *field*, *tenor* and *mode*. *Experiential Domain* in *Field* specifies the theme of the text whereas *Goal Orientation* denotes the function of the text. Here we distinguish between different functions: *exposition*, *narration*, *argumentation*, *persuasion*, and *instruction*. *Tenor* specifies further *Agentive Role*, *Social Role* and *Social Distance*. *Agentive Role* denotes the relation between the author and his addressee. Here the annotator can choose between: *expert to expert*, *expert to layperson*, *layperson to expert*, *layperson to layperson*. The *Social Role* denotes the position of the author as compared to his addressee: *equal* or *unequal*. The *Social Distance* can be: *casual*, *neutral*, *formal*, *intimate*, *colloquial*, and *consultative*. *Mode* specifies also three sub-categories: *Language Role*, *Channel* and *Medium*. *Language Role* can be: *ancillary* or *constitutive*. The *Channel* field states whether the text is *graphic* (printed) or *electronic*. *Medium* records if the text was found in a *written*, *spoken* or *written to be spoken* form.

Figure 12: CroCo-Meta for header annotation

The *Encoding Description* part (see Figure 13) of the CroCo-Meta form encodes information about how the corpus was built, the total number of words, if it was totally incorporated into the corpus or

⁷ An introduction to register analysis can be found in Ghadessy (1993).

only partially, when it was incorporated into the corpus and what language variation is used, for instance, British English or American English.

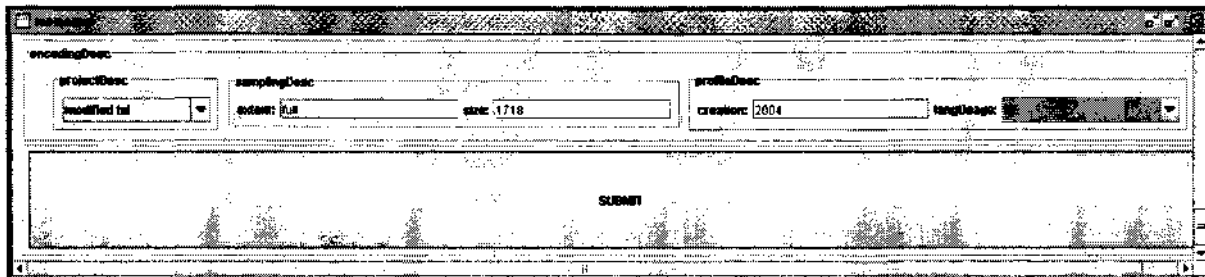
The image shows a web browser window displaying a form titled 'encodingDesc'. The form contains several input fields: 'projectDesc' with a dropdown menu showing 'modified.txt', 'samplingDesc' with a dropdown menu showing 'full', 'size' with the value '1718', 'profileDesc' with a dropdown menu showing '2004', and 'language' with a dropdown menu. Below these fields is a large 'SUBMIT' button.

Figure 13: Encoding Description in CroCo-Meta

As shown in Figure 14 the CroCo-Meta form can also be used to document the process of annotation. This kind of information is specified by the *Annotation Description* and *Alignment Description* part in the CroCo-Meta form. For each alignment or annotation level the form provides fields for entering the first annotator, the person who did the correction and consistency check of the annotation. Through the documentation of the annotation and alignment process problems and changes become more visible and can later be easily identified and tracked back.

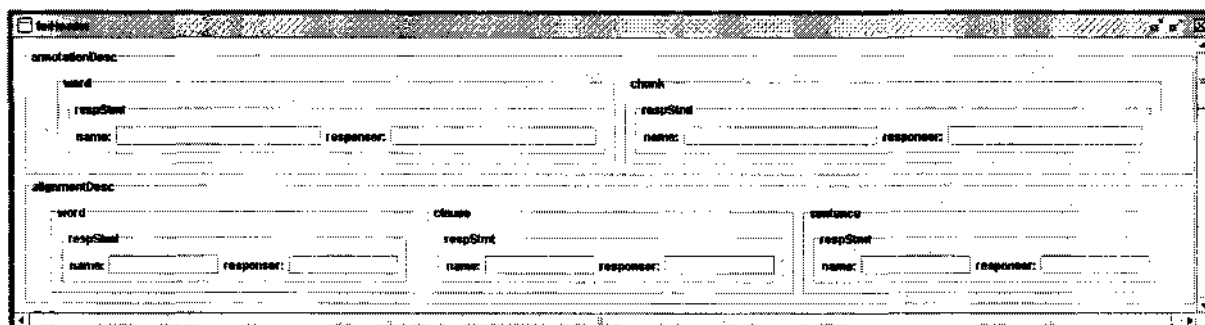
The image shows a web browser window displaying a form titled 'annotationDesc' and 'alignmentDesc'. The 'annotationDesc' section has two columns of fields. The first column has 'name:' and 'response:' fields. The second column has 'name:' and 'response:' fields. The 'alignmentDesc' section has three columns of fields. The first column has 'word', 'name:', and 'response:' fields. The second column has 'clause', 'name:', and 'response:' fields. The third column has 'sentence', 'name:', and 'response:' fields.

Figure 14: CroCo-Meta for the documentation of the annotation and alignment process

As described above the tool CroCo-Meta provides (for both Windows and Linux) an easy way for the storage of meta-information. Thus annotating the header is facilitated, since information is presented as fields of a form. This means that the annotator does not need to work with XML files in which multiple embedding of elements may occur. Furthermore, the presetting of the possible entry choices makes the annotation process faster and diminishes the error rate. Additionally, the storage of meta-information can be useful when it comes to query a specific register or specific text type (e.g. written text vs. spoken text). The tool itself can easily be adapted to other projects and is free for non-profit research purposes.

```

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE teiHeader SYSTEM "header.dtd">
<teiHeader>
<fileDesc>
<filename>GO_SHARE_001.txt</filename>
<subcorpus>SHARE GO</subcorpus>
<language>German</language>
<titleStmt>
<title>Jenoptik - Brief an die
Aktionäre (1999, 2000,
2001)</title>
<author>Spät, Lothar
(Vorstandsvorsitzender)</author>
</titleStmt>
<translation></translation>
<publicationStmt>
<publisher>Universität
Hamburg</publisher>
<date>1999-2001</date>
<distributor>www.rrz.uni-
hamburg.de/SFB538/forschung/
kommunikation/k4.html
(Verdecktes Übersetzen - TP K4)
</distributor>
<availability>local</availability>
</publicationStmt>
<sourceDesc>
<author/>
<title/>
</sourceDesc>
<registerAnalysis>
<register>SHARE</register>
<field>
<experientialDomain>report about
the growth and the goals of
Jenoptik, investments in the
future and the enhancement of the
company; Jenoptik thanks their
shareholders and their employees
for the good cooperation in the
last years</experientialDomain>
<goalOrientation>persuasion
</goalOrientation>
</field>
<tenor>
<agentiveRole>expert to
layperson</agentiveRole>
<socialRole>equal</socialRole>
<socialDistance>formal
</socialDistance>
</tenor>
<mode>
<languageRole>constitutive
</languageRole>
<channel>graphic</channel>
<medium>written</medium>
</mode>
</registerAnalysis>
</fileDesc>
<encodingDesc>
<projectDesc>modified
tei</projectDesc>
<samplingDesc>
<extent>full</extent>
<size>1.718</size>
</samplingDesc>
<profileDesc>
<creation>2004</creation>
<langUsage>DE</langUsage>
</profileDesc>
</encodingDesc>
<annotationDesc>
<alignment>
<respStmt>
<name/>
<responser/>
</respStmt>
</alignment>
<morph>
<respStmt>
<name/>
<responser/>
</respStmt>
</morph>
<chunk>
<respStmt>
<name/>
<responser/>
</respStmt>
</chunk>
</annotationDesc>
</teiHeader>

```

Figure 15: The header in XML format after the CroCo-Meta form was filled and submitted

6. Conclusions

The need for linguistically annotated corpora is observed across all branches of linguistics, and the translation branch is no exception. One of the reasons why big parallel corpora with high quality annotation and alignment are still not available for use in machine translation projects is due to the fact that collecting and building such a corpus is a time-consuming and expensive task.

There are certainly research questions which do not require such a detailed linguistic analysis. Quantitative automatic annotation allows to a certain point to resolve the research questions without the burden of using or constructing a big resource with qualitative enrichment in advance. However, more complex research questions, which deal with the specificity of a language need a more detailed linguistic analysis in order to be answered.

The research described here shows the use of linguistic annotated corpora across languages. We have shown that in order to solve typical translation problems for English and German, such as English cleft constructions, linguistic information facilitates this. Information about part-of-speech, grammatical function assignment, phrasal structure identification as well as alignment on different levels can help to identify the correct construction. This type of translation memory look-up enables the translator to search for more than one problematic lexico-grammatical construction and its aligned translation.

In section 2 we described the use of corpora in translation studies and argued for linguistic annotated data when it comes to translation. In section 3 we showed what kind of linguistic information can be encoded and stored and by what means. Section 4 shows how annotated data is used when it comes to query for complex grammatical constructions. Because certain grammatical properties are text-specific, section 5 introduces a tool for annotating meta-information and shows how such a tool can specialise the memory look-up, for instance, by searching for cleft constructions in a specific publication.

The MySQL storage of linguistically annotated data combined with the possibility to exploit this data easy and fast allows the extraction and comparison of grammatical complex structures across languages.

7. References

- Baker, M. 1996. Corpus-based translation studies: The challenges that lie ahead. In: Somers, H. L. (ed). *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam: Benjamins, 175-186.
- Baker, M. 2000. Towards a Methodology for Investigating the Style of a Literary Translator. In: *Target* (12)2, 241-266.
- Barlow, M. 2000. Parallel texts in language teaching. In: Botley, S.P., McEnery, A.M. & Wilson, A. (eds) *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 106-115.
- Brants, T. 2000. TnT - A Statistical Part-of-Speech Tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.
- Bernardini, S. and F. Zanettin. 2004. When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals. In: Kujamäki/Mauranen 2004, 51-62.
- Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8/4:243-257.
- Bowker, L. 1999. Exploring the potential of corpora for raising language awareness in student translators. In: *Language Awareness* (8), 160-173.
- Bowker, L. and P. Bennison. 2003. Student translation archive: Design, development and application. In: Zanettin/Bernardini/Stewart 2003, 103-117.
- Bowker, L. and J. Pearson. 2002. *Working with Specialized Language: A practical guide to using corpora*. London etc.: Routledge.
- Doherty, M. 1999. Clefts in translations between English and German. In: *Target* (11)2, 289-315.
- Fabricius-Hansen, C. 1999. Information packaging and translation: Aspects of translational sentence splitting (German - English/Norwegian). In: Doherty, M. (ed) *Sprachspezifische Aspekte der Informationsverteilung/(studia grammatica 47)* Berlin: Akademie-Verlag, 175-214.

- Ghadessy, Mohsen. 1993. *Register Analysis, Theory and Practise*. Continuum International Publishing Group.
- Hansen, Silvia. 2003. *The Nature of Translated Text. An interdisciplinary methodology for the investigation of the specific properties of translations*. Saarbrücken: Saarbrücken Dissertations in Computational Linguistics and Language Technology. Vol. 13.
- Hansen-Schirra, Silvia, and Stella Neumann and Mihaela Vela, 2006. Multi-dimensional Annotation and Alignment in an English-German Translation Corpus. In *Proceedings of the workshop on NLPXML-2006*. Italy.
- Hansen, S. and Elke Teich. 1999. Kontrastive Analyse von Übersetzungskorpora: ein funktionales Modell. In: Gippert, J. (ed) *Sammelband der Jahrestagung der GLDV99*. Prag: enigma corporation, 311--322.
- Hansen, S. and Elke Teich. 2001. Multi-layer analysis of translation corpora: methodological issues and practical implications. In: *Proceedings of EUROLAN 2001 Workshop on Multi-layer Corpus-based Analysis*. Iasi, 44-55.
- Hasselgård, H. 1998. Thematic structure in translation between English and Norwegian. In: Johansson, S. & Oksefjell, S. (eds) *Corpora and Cross-linguistic Research: Theory, Method and Case Studies*. Amsterdam: Rodopi.
- Hawkins, John. 1986. *A comparative typology of English and German. Unifying the contrasts*. London, Croom Helm.
- Heyn, Matthias. 1996. Integrating machine translation into translation memory systems. *European Association for Machine Translation - Workshop Proceedings, ISSCO, Geneva*: 111-123.
- Johansson, S. 1995. Mens sana in corpora sano: On the role of corpora in linguistic research. In: *The European English Messenger* (4)2, 19-25.
- Johansson, S. 2003. Reflections on corpora and their uses in cross-linguistic research. In: Zanettin/Bernardini/Stewart 2003, 133-144.
- Johansson, S. and K. Hofland. 2000. The English-Norwegian Parallel Corpus: Current work and new directions. In: Botley, S.P., McEnery, A.M. & Wilson, A. (eds.) *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi.
- Kenny, D. 1998. Creatures of Habit? What translators usually do with words. In: *Meta: Special Issue on The Corpus-Based Approach* 43(4), 515-523.
- Kilgarriff, A. 2001. Web as corpus. In: *Proceedings of Corpus Linguistics 2001*. Lancaster, 342-344.
- Kujamäki, P. and A. Mauranen. 2004. *Translation Universals: Do They Exist?* Amsterdam: Benjamins.
- Laviosa-Braithwaite, S. (1996). *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. (PhD Thesis). Manchester: UMIST.
- Maas, Heinz Dieter. 1998. Multilingual Textverarbeitung mit MPRO. *Europäische Kommunikationskybernetik heute und morgen '98*, Paderborn.

- Maia, B. 2003. Some languages are more equal than others: training translators in terminology and information retrieval using comparable and parallel corpora. In: Zanettin/Bernardini/Stewart 2003, 43-53.
- Malmkjær, K. 2003. On a pseudo-subversive use of corpora in translator training. In: Zanettin/Bernardini/Stewart 2003, 119-134.
- Mauranen, A. 1997. Hedging in language revisers' hands. In: Markkanen, R. & Schröder, H. (eds) *Hedging and Discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts*. Berlin: Walter de Gruyter, 115-133.
- Sperberg-McQueen, Christopher Michael & Lou Burnard (eds.), 1994. *Guidelines for Electronic Text Encoding and Interchange (TEIP3)*. Text Encoding Initiative, Chicago and Oxford.
<http://www.tei-c.org/>
- Müller, Christoph and Michael Strube. 2003. Multi-Level Annotation in MMAX. *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan: 198-107.
- Neumann, Stella and Silvia Hansen-Schirra. 2005. The CroCo Project: Cross-linguistic corpora for the investigation of explicitation in translations. In *Proceedings from the Corpus Linguistics Conference Series*, Vol. 1, no. 1, ISSN 1747-9398.
- Olohan, Maeve. 2003. How frequent are the contractions? A study of contracted forms in the Translational English Corpus. In: *Target* (15), 59-89.
- Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. London etc.: Routledge.
- Olohan, Maeve and Mona Baker. 2000. Reporting *that* in Translated English. Evidence for Subconscious Processes of Explicitation? *Across Languages and Cultures* 1(2):141-158.
- Pearson, J. 2000. Teaching terminology using electronic resources. In: Botley, S.P., McEnery, A.M. & Wilson, A. (eds) *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi, 92-115.
- Pearson, J. (2003), Using parallel texts in the translator training environment. In: Zanettin/Bernardini/Stewart 2003, 15-24.
- Teubert, W. 2001. *Text Corpora and Multilingual Lexicography: Special Issue of International Journal of Corpus Linguistics*.
- Geoffrey Sampson. 1995. English for the Computer. The Susanne Corpus and Analytic Scheme. Clarendon Press, Oxford.
- Schiller, Anne and Simone Teufel and Christine Stockert. 1999. *Guidelines für das Tagging deutscher Textkorpora mit STTS*, University of Stuttgart and Seminar für Sprachwissenschaft, University of Tübingen.
- Schrader, Bettina. 2006. ATLAS — a new text alignment architecture. In: *Proceedings of the Joint Coling/ACL Conference*, July 2006, Sydney, Australia
- Steiner, E. 2002. Grammatical metaphor in translation — some methods for corpus-based investigations. In: Behrens, B., Fabricius-Hansen, C., Hasselgård, H. & Johansson, S. (eds) *Information Structure in a Cross-linguistic Perspective*. Amsterdam: Rodopi.

- Teich, E. and Silvia Hansen. 2001a. Methods and techniques for a multi-level analysis of multilingual corpora. In: *Proceedings of Corpus Linguistics 2001*. Lancaster, 572-580.
- Teich, E. and Silvia Hansen. 2001b. Towards an integrated representation of multiple layers of linguistic annotation in multilingual corpora. In: *Online Proceedings of Computing Arts 2001: Digital Resources for Research in the Humanities*. Sydney.
- Teich, E. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin: Walter de Gruyter.
- Vintar, Spela and Silvia Hansen. 2005. Cognates — Free Rides, False Friends or Stylistic Devices: A Corpus-Based Comparative Study. In: Barnbrook, G., Danielsson, P. & Mahlberg, M. (eds) *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Birmingham: Birmingham University Press.
- Zanettin, F. and S. Bernardini and D. Stewart. 2003. *Corpora in Translator Education*. Manchester: St Jerome.