Querying Multi-Layer Annotation and Alignment in Translation Corpora

Mihaela Vela,[1] Stella Neumann[2]
and Silvia Hansen-Schirra[3]

1. Introduction

When dealing with linguistically annotated and aligned corpora current research concentrates mainly on the investigation of translation properties. However, annotated and aligned corpora can be useful for practical translation as well, since translators also work with parallel corpora. Translators typically use raw sentence aligned corpora stored in translation memories. In this paper we will show how linguistically annotated and aligned corpora can be exploited for both, research topics and practical applications when it comes to typological differences between languages and register-specific language use. In this context we present two different query catalogues: an application-oriented set of queries for computer-assisted translation and a research-oriented set of queries aiming at register analysis.

For this purpose we first present the corpus under investigation, its design, annotation and alignment in section 2. In this section we also introduce the MySQL database created for querying the annotated and aligned corpus. The query catalogues presented in section 3 and 4 offer a practical grammar-oriented look-up to be used in translator training and practice as well as rich linguistics-based insights into peculiarities of original registers and their translations. Section 5 briefly summarizes the findings of the paper and gives an outlook on future work.

2. The CROCO corpus

For the two types of queries described in this paper we use the CroCo Corpus (*cf.* Hansen-Schirra *et al.*, 2006), a linguistically annotated and aligned corpus of 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 the eight registers political essays (henceforth ESSAY), fictional texts (FICTION), instruction manuals (INSTR), popular scientific texts (POPSCI), shareholder information (SHARE), prepared speeches (SPEECH), tourism texts (TOU) and websites (WEB). 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 approximately one million words. Additionally, reference corpora are included for German and English. These register-neutral corpora (see Figure 1) include 2,000 word samples from 17 registers (see Neumann and Hansen-Schirra, 2005 for more details on the CroCo corpus design).

[1] Applied Linguistics, Translation and Interpreting, Saarland University, Saarbrücken, Germany
e-mail: m.vela@mx.uni-saarland.de
[2] Applied Linguistics, Translation and Interpreting, Saarland University, Saarbrücken, Germany
e-mail: st.neumann@mx.uni-saarland.de
[3] Applied Linguistics, Johannes Gutenberg University Mainz, Germersheim, Germany
e-mail: hansenss@uni-mainz.de

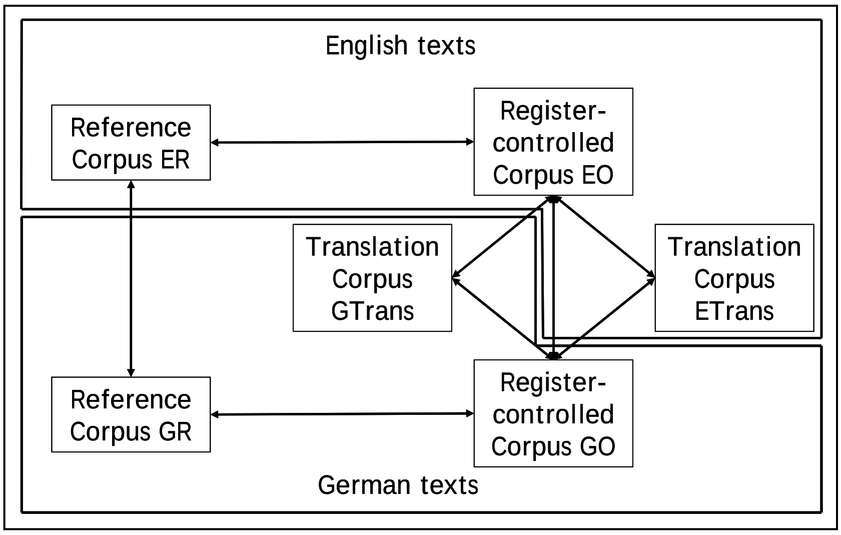

English texts
Reference Corpus ER
Register-controlled Corpus EO
Translation Corpus GTrans
Translation Corpus ETrans
Reference Corpus GR
Register-controlled Corpus GO
German texts

Figure 1: Corpus design in CroCo

The CroCo Corpus is tokenized 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 more detail in section 2.1. The transformation of the annotation and alignment into a MySQL database is described in section 2.2.

2.1 Multi-layer annotation and alignment

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 presented in this paper.

For each text in the corpus the annotation covers different levels. 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 following the TEI standard (Sperberg-McQueen and Burnard, 1994) in separate header files for each text in the corpus.

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 XCES[4]. 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 2 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.

[4] 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 2: Tokenisation and indexing

The second annotation layer is the part-of-speech layer, which is provided again by TnT[5]. 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 3. 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="mc1" 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 3: PoS tagging

Morphological information is particularly relevant for German due to the fact that this language carries grammatical 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 works analogous to the part-of-speech encoding shown in Figure 3.

Moving up from the token unit to the chunk unit, the same pattern as for the tokens is repeated. The annotation on the chunk level covers both, grammatical functions and phrase

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

category for the highest node in the sentence. This information is manually annotated with MMAX2 (Müller and Strube, 2003). MMAX2 is a customizable tool for creating annotations on multiple levels. One of its advantages is that it allows discontinuous units and links between units (relevant for our manual clause alignment, see below).

As for tokens, the chunks are also indexed each chunk having an **id** attribute as shown in Figure 4.

```
<?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 4: Chunk indexing

The phrase structure annotation (see Figure 5) assigns the **type** attribute to each phrase chunk identified in the manual chunk annotation. XPointers link the phrase structure annotation to the chunk index file.

```
<?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 5: Phrase structure annotation

The (manual) annotation of grammatical functions is again kept in a separate file and is comparable to the phrase structure annotation.

On clause and sentence level the information refers to the segmentation of the text in clauses and sentences. As shown in Figure 6, each clause consists of a list of XLinks to tokens in the index file denoted by the **xml:base** attribute. The same approach also applies 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="cl1">
   <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="cl2">
   <tok xlink:href="#t21"/>
   <tok xlink:href="#t22"/>
   <tok xlink:href="#t23"/>
   <tok xlink:href="#t24"/>
   <tok xlink:href="#t25"/>
  </clause>
  …
  <clause id="cl17">
   <tok xlink:href="#t168"/>
   <tok xlink:href="#t169"/>
   <tok xlink:href="#t170"/>
   <tok xlink:href="#t171"/>
   <tok xlink:href="#t172"/>
  </clause>
 </clauses>
</document>
```

Figure 6: 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 GIZA++ (Och and Ney, 2003) a statistical alignment tool. Clauses are aligned manually again with the help of MMAX 2. Finally, sentences are aligned using WinAlign, 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 the XCES standard.

Figure **7** 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. In the example given in Figure

7, there is no alignment for clause 2 (“#cl2”), resulting in an “undefined” tag for the translation.

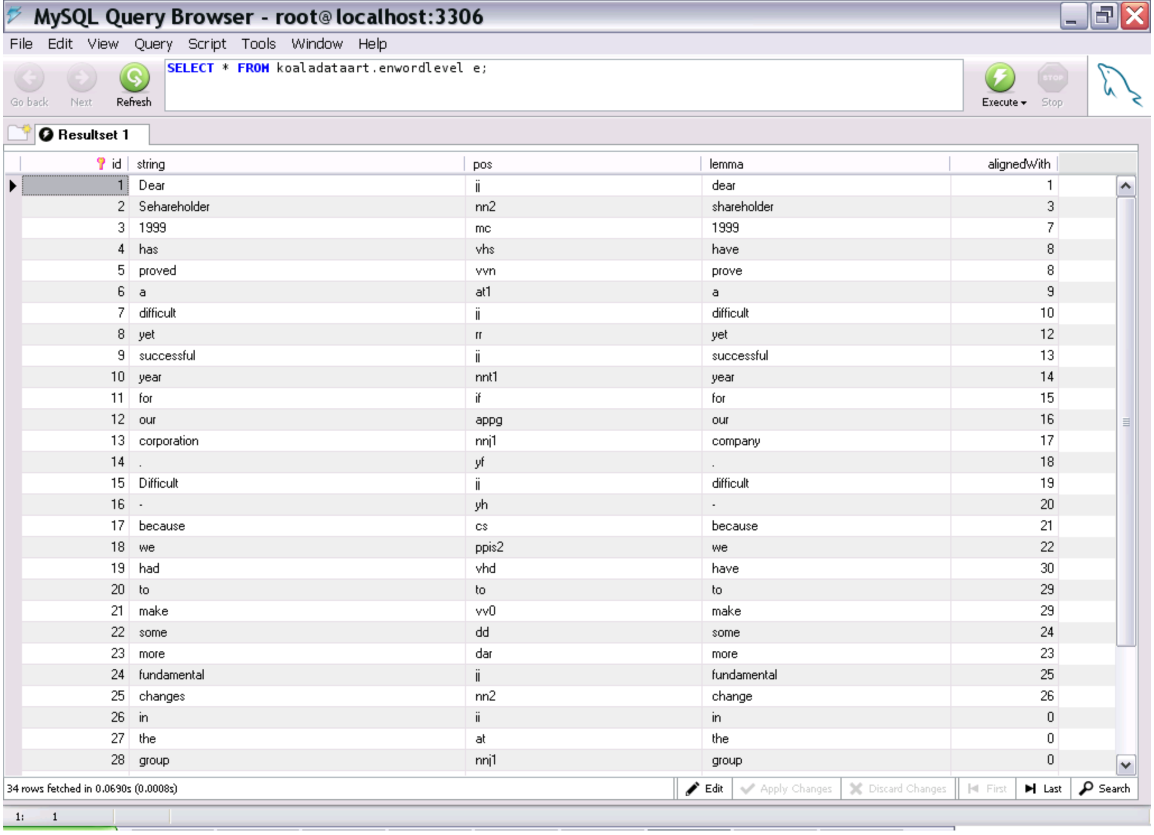
```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE document SYSTEM "clauseAlign.dtd">
<document xmlns:xlink="http://www.w3.org/1999/xlink"
 name="E2G_SHARE_001.clauseAlign.xml">
 <translations
  xml:base="CROCO_CORPUS/ENGLISH2GERMAN/GTrans/SHARE/ANNOTATED/clause/">
  <translation trans.loc="EO_SHARE_001.clause.xml" xml:lang="en" n="1"/>
  <translation trans.loc="GTrans_SHARE_001.clause.xml" xml:lang="ge" n="2"/>
 </translations>
 <clauses>
  <clause>
   <align xlink:href="#cl1"/>
   <align xlink:href="#cl1"/>
  </clause>
  <clause>
   <align xlink:href="#cl2"/>
   <align xlink:href="#undefined"/>
  </clause>
  <clause>
   <align xlink:href="#cl3"/>
   <align xlink:href="#cl3"/>
  </clause>
  <clause>
   <align xlink:href="#cl4"/>
   <align xlink:href="#cl4"/>
  </clause
 …
 </clauses>
</document>
```

Figure 7: Alignment of clauses

The alignment of tokens and sentences in the corpus follows the same principle. Additionally, phrase alignment can be derived from the word alignment in combination with the phrase structure annotation, and grammatical functions can be mapped automatically across the parallel corpus.

2.2 The database

The annotation and alignment described in the previous section is the basis for experiments facilitating the query of linguistic information in translation corpora. In order to achieve a fast and efficient search we converted the annotation and alignment presented above into tables of a MySQL database.



MySQL Query Browser - root@localhost:3306
File Edit View Query Script Tools Window Help
Go back
Next
Refresh
Execute
Stop
Resultset 1

```
SELECT * FROM koaladataart.enwordlevel e;
```

id	string	pos	lemma	alignedWith
1	Dear	jj	dear	1
2	Sehareholder	nn2	shareholder	3
3	1999	mc	1999	7
4	has	vhs	have	8
5	proved	vvn	prove	8
6	a	at1	a	9
7	difficult	jj	difficult	10
8	yet	rr	yet	12
9	successful	jj	successful	13
10	year	nnt1	year	14
11	for	if	for	15
12	our	appg	our	16
13	corporation	nnj1	company	17
14	.	yf	.	18
15	Difficult	jj	difficult	19
16	-	yh	-	20
17	because	cs	because	21
18	we	ppis2	we	22
19	had	vhd	have	30
20	to	to	to	29
21	make	vv0	make	29
22	some	dd	some	24
23	more	dar	more	23
24	fundamental	jj	fundamental	25
25	changes	nn2	change	26
26	in	ii	in	0
27	the	at	the	0
28	group	nnj1	group	0

34 rows fetched in 0.0690s (0.0008s)

Figure 8: 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 one set of tables in the database (see Figure 8). The English tokens in Figure 8 are indexed, each index being assigned a string, a lemma, a part-of-speech tag and an index for its German equivalent. At chunk level, a set of tables is filled with information about chunk type and the grammatical function the chunk fulfils. Similarly to the XML encoding of the corpus, the MySQL tables for chunks are 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 gives us an easier and faster 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.

In the following two sections we will discuss possible queries from two perspectives. In section 3, we will present queries aimed at the practical translator. In section 4, we will exemplify queries relevant for linguistic research.

3. Application-oriented queries

In many cases, typological differences between languages do not pose any problems in the translation process. Different word order or grammatical morphologies are not considered as major translation problems. There are, however, typological differences that are problematic for the translator. Typically, these are grammatical 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. However, it is not always easy to find an adequate translation equivalent. For this reason, a query catalogue and the resulting translation examples of handling typological differences can help to solve translation problems.

In the following, we explain the advantage of such a catalogue on the basis of Hawkins' (1986) account of systematic contrastive differences between English and German applied to the database described in section 2. Among others, Hawkins states that English is far more productive concerning raising constructions, cleft sentences and deletions. Therefore, in the process of translating these constructions into English, compensations have to be found.

One of the constructions for which English is more productive than German is the raising construction. It can be found by querying the following pattern in plain text:

> grammatical function=”finite verb” (FOLLOWED BY grammatical 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:

(1) We continue to benefit from the strong natural gas market in North America. --- Wir profitieren weiterhin von einem starken Erdgasmarkt in Nordamerika.

(2) We defined the minivan, and will continue to do so. --- Wir haben den Minivan erfunden und wir werden auch künftig neue Marktsegmente definieren.

(3) … 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 strategy can be identified. The meaning of the verb “continue” which occurs frequently in the English SHARE sub-corpus is translated by using temporal adverbials in German (“weiterhin” and “künftig” in (1) and (2)). Additionally, in (3) the verbal group is transformed into a nominal structure which seems to be a typical translation strategy for the translation direction English-German.

Cleft constructions are another typical feature of the English grammatical system. While they do exist in German as well, German frequently uses other options of realizing information distribution patterns, e.g., 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 grammatical function=”complement” (INCLUDING part-of-speech=”relative pronoun”))

Applying this query to our database, we find the following translation pairs[6]:

(4) 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.

(5) 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.

(6) 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.

The examples illustrate two options of translating English clefts into German. (4) is realised in the German version by a prepositional phrase, whereas in (5) and (6) 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” (English gloss: “it is somebody/something to thank that”) for the translation of both English cleft sentences. This

[6] These examples are taken from our English-German SHARE sub-corpus.

might be an indicator for a good translation strategy 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.

According to Hawkins, substitutions and deletions occur more frequently in English than in German. We search for deletions using the following patterns. The query described below retrieves all nominal phrases in which the nominal head is deleted.

phrasal category=”prepositional phrase / noun phrase” NOT INCLUDING part-of-speech=”noun”

Surprisingly, we found more deletions in the German translations than in the English originals. Example (7) shows a nominal substitution in the English original translated by a German prepositional phrase in which the nominal head is deleted. Since substitution is not as readily available in German, deletion seems to be one strategy to translate this structure.

(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.

Coordinated sentences with one common subject are retrieved by the following query.

phrasal category=”sentence” INCLUDING 2 * grammatical function=”finite verb”
AND 1* grammatical function=”subject”

In (8), the subject “we” in the English original is repeated in the second clause featuring a new verb. In the translation, we again find two coordinated clauses with two different German verbs (“danken” and “tun”). However, here the subject is deleted in the second clause. The same phenomenon can be observed in (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.

(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.

Another application of our word alignment is its use to create a bilingual dictionary or a bilingual term base. We can extract, for example, specific categories (like verbs and nouns) from our aligned corpus. Other combinations are also possible: For instance, German tends to use many compounds, which typically correspond to multiword units in English. Such multiword alignments can be extracted automatically (cf. Schrader, 2006) from the corpus[7] (see Figure 9).

```
<item>
   <lemma>silk handkerchief</lemma>
   <category>multiword</category>
```

[7] The following to examples are taken from our FICTION sub-corpus.

```
   <language>English</language>
 <translations>
  <translation>
   <lemma>einstecktuch</lemma>
   <category>noun</category>
   <language>German</language>
  </translation>
 </translations></item>
```

Figure 9: English-German word alignment displaying multiword units

Translation shifts are another issue of interest for the practical translator, for instance transpositions (shifts in word class, *cf.* Vinay and Darbelnet, 1995). In order to detect these, verb-noun alignments can be extracted as displayed in Figure 10.

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

Figure 10: English-German word alignment displaying a transposition

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

Beyond their use for the practitioner, these queries are relevant to the scholar as well. From the research point of view a whole range of queries is of interest. These queries will be discussed in the following section going into some detail with regard to the interpretation of the findings.

4. Research queries

The research-oriented query catalogue is based on register analysis (for a basic account of register analysis *cf.* Halliday and Hasan, 1989). Register analysis investigates referential information (under the heading “field of discourse”), pragmatic aspects (“tenor of discourse”) and the means used to create a textual whole (“mode of discourse”). These general parameters have been further subdivided into various subdimensions which in turn are finally related to operationalisations, i.e. observable linguistic indicators in texts (*cf.* Steiner, 2004, Hansen-Schirra *et al.*, to appear). Register analysis can be used as a comprehensive set of criteria for the analysis of individual texts. It can also be employed to investigate corpora containing large amounts of texts from different registers in order to identify variation within a given register, across registers and between originals and translations.

While the annotation and alignment of the CroCo Corpus as presented in section 2 offers a wide range of linguistic information, quantitative queries of the type discussed here necessarily result in a reduction in informativity as compared to qualitative example-based analyses. Conversely, the former allow statements on the relevance of the investigated features for a given register.

In the following subsections we will present a selection of queries contributing to a comprehensive register analysis organized by the three main parameters of field, tenor and mode of discourse. Space does not allow covering all features in the query catalogue, but we discuss a range of features of different degrees of difficulty and complexity to demonstrate the feasibility of the query catalogue.

4.1 Field of discourse

Under the parameter “field of discourse” we are mainly interested in two dimensions, experiential domain, i.e. the area of referential meaning covered by the register, and goal orientation, i.e. the aims typically pursued by the authors of texts belonging to the given register.

The most frequent vocabulary gives us an important indication of the experiential domain covered by the register. It is straightforward to query vocabulary by running frequency word lists on each individual text with the help of a concordance tool like WordSmith (Scott, 1996). Beyond the mere frequency of the respective lemma in a word list lexical chains, i.e. sequences of related words (Morris and Hirst, 1991), provide information on whether a frequent lemma forms a topical thread throughout the text or whether it is repeated only locally. In the former case the chain underpins the lemma’s relevance for determining the referential meaning of the whole text. In the latter case the lemma forming a chain represents only a local strand of referential meaning.

For a comprehensive analysis of lexical chains in each text, the corpus has to be annotated with sense relations, since semantically related items like synonyms, hyponyms *etc.* must be interpreted as contributing to a lexical chain (e.g. with the help of a WordNet and GermaNet annotation which is currently not available for the CroCo Corpus). The present account thus concentrates on chains created by repetitions of the same lemma.

The MYSQL-based query looks for each lemma and the sentence ID in which the given lemma appears. Since the sentence IDs correspond to a sentence’s linear position in a text, the IDs of two consecutive occurrences represent the distance between the two links in the chain. Morris and Hirst (1991) also mention the span from the first to the last occurrence of the lemma within the text. This span is an additional cue to the relevance of the chain for the overall referential meaning of the text. We interpret the repetition of the lemma as a link in a continuous chain if the distance between the occurrences is less than four sentences. If the distance is longer, the new occurrence of the lemma is interpreted as a return to an existing chain (Morris and Hirst, 1991: 32). Chain lengths thus result from the addition of occurrences in sentences less than 4 sentences away. Table 1 displays extracts from the most frequent content word in two text pairs from the SHARE and the FICTION sub-corpora, spans and distances between each occurrence in sentences as well as the resulting chain lengths.

<table>
<tr><th colspan="24">EO_SHARE_004: "GE"</th></tr>
<tr><td><b>Text length in sentences</b></td><td colspan="23">166</td></tr>
<tr><td><b>lemma frequency</b></td><td colspan="23">56</td></tr>
<tr><td><b>span</b></td><td colspan="23">161</td></tr>
<tr><td><b>distance of occ.</b></td><td>2</td><td>6</td><td>2</td><td>2</td><td>0</td><td>1</td><td>1</td><td>1</td><td>1</td><td>6</td><td>1</td><td>1</td><td>1</td><td>2</td><td>8</td><td>2</td><td>1</td><td>1</td><td>28</td><td>2</td><td>0</td><td>1</td><td>5</td></tr>
<tr><td><b>chain length</b></td><td>1</td><td colspan="8">8</td><td colspan="5">5</td><td colspan="4">4</td><td colspan="4">4</td><td>10</td></tr>
</table>

<table>
<tr><th colspan="24">GTrans_SHARE_004: "GE"</th></tr>
<tr><td><b>Text length in sentences</b></td><td colspan="23">166</td></tr>
<tr><td><b>lemma frequency</b></td><td colspan="23">48</td></tr>
<tr><td><b>span</b></td><td colspan="23">157</td></tr>
<tr><td><b>distance of occ.</b></td><td>6</td><td>2</td><td>1</td><td>1</td><td>0</td><td>1</td><td>1</td><td>1</td><td>1</td><td>4</td><td>2</td><td>1</td><td>1</td><td>1</td><td>6</td><td>6</td><td>1</td><td>1</td><td>26</td><td>1</td><td>1</td><td>2</td><td>0</td></tr>
<tr><td><b>chain length</b></td><td colspan="9">9</td><td colspan="5">5</td><td>1</td><td colspan="3">3</td><td colspan="5">6</td></tr>
</table>

<table>
<tr><th colspan="24">EO_FICTION_001: "day"</th></tr>
<tr><td><b>Text length in sentences</b></td><td colspan="23">207</td></tr>
<tr><td><b>lemma frequency</b></td><td colspan="23">25</td></tr>
<tr><td><b>span</b></td><td colspan="23">147</td></tr>
<tr><td><b>distance of occ.</b></td><td>2</td><td>2</td><td>1</td><td>5</td><td>3</td><td>3</td><td>5</td><td>1</td><td>2</td><td>2</td><td>1</td><td>0</td><td>10</td><td>9</td><td>0</td><td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>1</td><td>18</td><td>33</td></tr>
<tr><td><b>chain length</b></td><td colspan="3">3</td><td colspan="3">3</td><td colspan="6">6</td><td>1</td><td colspan="8">8</td><td>1</td><td>1</td></tr>
</table>

<table>
<tr><th colspan="24">GTrans_FICTION_001: "Tag"</th></tr>
<tr><td><b>Text length in sentences</b></td><td colspan="23">208</td></tr>
<tr><td><b>lemma frequency</b></td><td colspan="23">21</td></tr>
<tr><td><b>span</b></td><td colspan="23">148</td></tr>
<tr><td><b>distance of occ.</b></td><td>2</td><td>2</td><td>1</td><td>5</td><td>5</td><td>5</td><td>1</td><td>2</td><td>2</td><td>1</td><td>0</td><td>10</td><td>9</td><td>0</td><td>2</td><td>1</td><td>18</td><td>10</td><td>24</td><td>48</td></tr>
<tr><td><b>chain length</b></td><td colspan="3">3</td><td>1</td><td>1</td><td colspan="6">6</td><td>1</td><td colspan="4">4</td><td>1</td><td>1</td><td>1</td><td>1</td></tr>
</table>

Table 1: Extracts from lexical chains per text

We can gather the following information from the Table 1:

- The most frequent content word in the FICTION text occurs less frequently than the corresponding word in the SHARE text although the text is longer (by 40 sentences), thus producing weaker lexical chains.
- The most frequent word in the original shareholder information text establishes longer chains than its counterpart in the original fictional text.
- Although it spans almost the whole text, the most frequent word in the SHARE texts has a break of 28 sentences in the original and 26 in the translation.
- The lexical chains in the translations are shorter in both registers.
- The spans in SHARE cover almost the complete text. The span in FICTION is clearly shorter.

From the very restricted example presented here we can assume that the referential meaning in the fictional text is more diffused compared to the SHARE text. This procedure repeated for all texts in the register results in information how much individual lexical items contribute to the establishment of the text's referential meaning. The break in the chain in the SHARE text points to a different local strand of meaning in this part of the text.

Assessing reiteration, i.e. repetitions within a chain, density, i.e. proximity of the chain links, and length of the chains (cf. Morris and Hirst, 1991: 32), will reveal registerial differences and similarities. The inclusion of translated texts, as in the example discussed here, will help determining shifts occurring in the translation process.

Goal orientation refers to the aims authors pursue with their texts in terms of generic types like exposition, instruction, narration, argumentation, persuasion *etc.* The features analyzed under this heading therefore have to be interpreted with respect to these types. The features comprise modality, mood, pronominalisation, rhetorical style, tense, theme/rheme and voice *etc.* It is not possible to illustrate the complete interplay of these features for the interpretation of generic types in this paper. What we can do is exemplifying the analysis and interpretation of one feature. We have picked out past tense as an indicator relevant for the interpretation of generic types.

Past tense can be retrieved by querying the morphological annotation in CroCo.[8] More specifically, the query retrieves the value **pret** for English and **past** for German for the attribute **tns** in the morphology annotation. Figure 11 shows the distribution of past tense in all eight CroCo original registers grouped by languages. The number of all finite verbs serves as the basis of comparison and the results are given as the difference in percentage points between the register and the respective reference corpus. That is, if the value in the figure is negative, the respective register contains fewer verbs in the past tense than the reference corpus. This interpretation is inspired by Biber's work including not only features typical for a given group of texts but also features whose relative absence is characteristic for these texts (*cf.* Biber, 1995).

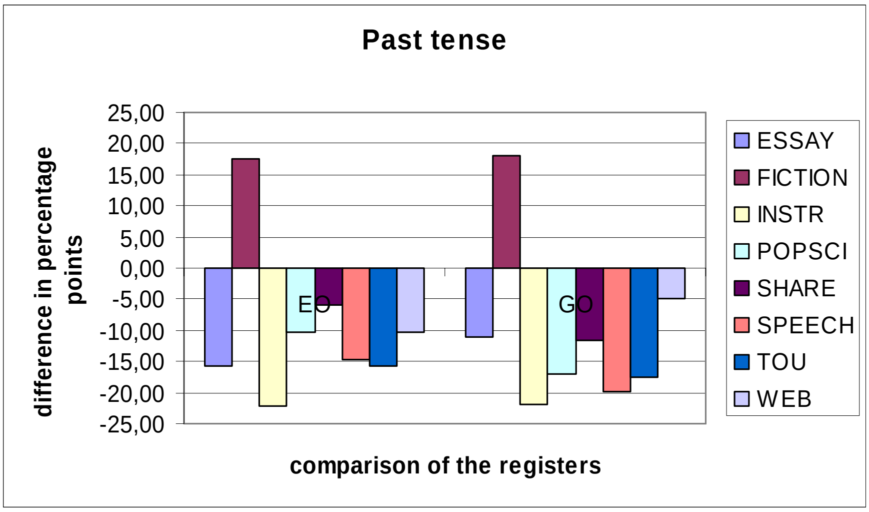


Figure 11: Difference of past tense values between original registers and reference corpora

FICTION clearly stands out in both languages with significantly higher values for past tense as compared to the reference corpora. All other CroCo registers exhibit clearly lower values than the reference corpora. Instructional texts contain the least verbs in past tense in both languages. Concentrating on these two registers we can attempt the following interpretation in view of goal orientation. Past tense is widely described as a typical feature of narrative genres (*cf.*, for instance, Santini, 2006 with further references). We can thus assume that fictional texts – not surprisingly - carry indicators of narration. This may seem a blunt statement, but the findings presented here confirm this assumption on a quantitative basis.

Just as past tense is a positive feature of narrative genres it seems to be a negative feature of instructional texts. The relative absence of past tense in instructional texts can be

[8] Unlike the two part-of-speech tag sets the morphology annotation scheme includes comparable tags for past tense in both languages.

explained by the procedural character of this genre focussing on the account of a sequence of procedures to be completed by the user of the device described in the manual. Past tense does not lend itself easily to this description.

The combination of the findings for "past tense" with findings for other features contributing to the interpretation of generic types will ultimately result in the description of realizational patterns typical for each type. It can be assumed that these patterns may be realized in a different way in translations thus diluting the generic types. This can be verified on the basis of the complete set of queries sketched here.

4.2 Tenor of discourse

Tenor of discourse is used to model the relationship between sender and recipient of a text. Among the dimensions to be described here is social hierarchy, sometimes also referred to as social role relationship. Its purpose is to determine whether sender and recipient hold equal social roles. These roles should be reflected in the linguistics choices the interactants make. Social hierarchy is further differentiated into level of authority, level of expertise, level of education. Other aspects contributing to an individual's position in the social hierarchy are religion, gender, sexual orientation *etc.*

We will demonstrate the query catalogue for level of expertise. Observable indicators are features of language for specific purposes (LSP) like LSP terminology and LSP grammar. While terminology can be detected again with the help of a concordance tool, e.g. with a key word analysis in WordSmith (Scott, 1996), LSP grammar can be queried on the basis of the CroCo annotation using the phrase chunking as well as sentence and clause segmentation. Grammatical structures typical for LSP texts have been described as packing more information into noun phrases and at the same time reducing the complexity of the clause structure (*cf.* Halliday and Martin, 1993).

SHARE	GO	ETrans	Diff.	EO	GTrans	Diff.
no. of texts	11	11	-	13	13	-
no. of sentences	1734	1738	4	1489	1467	-22
no. of clauses	2931	3797	866	3649	3097	-552
no. of chunks	9353	8602	-751	7251	8400	1149
no. of words	35223	39493	4270	35814	36370	556
chunks per sentence (av.)	5,39	4,95	-0,44	4,87	5,73	0,86
chunks per clause (av.)	3,19	2,27	-0,93	1,99	2,71	0,73
clauses per sentence (av.)	1,69	2,18	0,49	2,45	2,11	-0,34
words per sentence (av.)	20,31	22,72	2,41	24,05	24,79	0,74
words per clause (av.)	12,02	10,40	-1,62	9,81	11,74	1,93
words per chunk (av.)	3,77	4,59	0,83	4,94	4,33	-0,61
sentences per text (av.)	157,64	158,00	0,36	114,54	112,85	-1,69
clauses per text (av.)	266,45	345,18	78,73	280,69	238,23	-42,46
chunks per text (av.)	850,27	782,00	-68,27	557,77	646,15	88,38
words per text (av.)	3202,09	3590,27	388,18	2754,92	2797,69	42,77

Table 2: LSP grammar for the SHARE sub-corpus

Our query retrieves the number of phrases per clause and sentence, the number of words per sentence, clause and phrase *etc.* (see Table 2). Since the manual chunk annotation is still in progress we currently can only demonstrate this query for the SHARE sub-corpus. Differences between the registers can be determined once the query can be run over the whole corpus. The interpretation will go along the following lines. In LSP registers we would expect fewer clauses and more chunks per sentence, reflecting the tendency to pack information into nominal phrases rather than spread it over clauses. Consistently with this expectation, we would also expect a higher number of words per chunk.

For the time being we can interpret the findings presented in Table 2 in view of the difference between originals and translations. The differences between the values for originals and translations clearly point in one direction: Where the translations into English display a higher value, e.g. in the number of clauses (866 more than the German originals) the translations into German display a lower value (552 fewer clauses than the English originals). The units within the sentence (grouped here under the heading ‘LSP grammar’) thus seem to represent clear language-specific patterns. The translations therefore follow the typical structures of the respective target language. This finding can be interpreted in terms of normalization, i.e. the tendency of translations to adhere to or even exaggerate target language norms (*cf.* Baker, 1996).

4.3 Mode of discourse

This parameter covers all textual features giving information on the specific textual characteristics that help distinguish registers e.g. in terms of the role language plays for the constitution of a text. One of the features used as an indicator for medium, i.e. the dimension covering the spoken-written continuum, is lexical density. This ratio compares the number of content words with the number of all running words in a text (*cf.* Ure, 1971). It gives an indication of how much lexical content is spread over how many words under the assumption that registers typical for spoken interaction exhibit a lower lexical density (and at the same time more grammatical intricacy, *cf.* Halliday, 1989, and Ventola, 1996 with regard to academic writing).

For computing lexical density, we first have to retrieve the number of content words in a text. The query counts the lemmas (as part of the morphology annotation) of all words tagged with open class part-of-speech tags in the part-of-speech annotation (nouns, main verbs, adjectives, adverbs *etc.*). The number of all words per text can be obtained from the token index file.

Before discussing some results from the CroCo corpus, it should be noted that language typological differences between English and German in terms of the role function words play forbid direct comparisons of the lexical density ratios in the two languages (*cf.* Steiner, to appear for a more detailed discussion). Therefore our account concentrates on a comparison of the variation of the ratios in the five English original registers FICTION, INSTR, POPSCI, SHARE and SPEECH. **Figure 12** presents a box plot from the SPSS descriptive statistics displaying information like median, interquartile range, i.e. the spread of the middle 50 percent of the scores represented by the boxes, and outliers (marked with the number of the case).

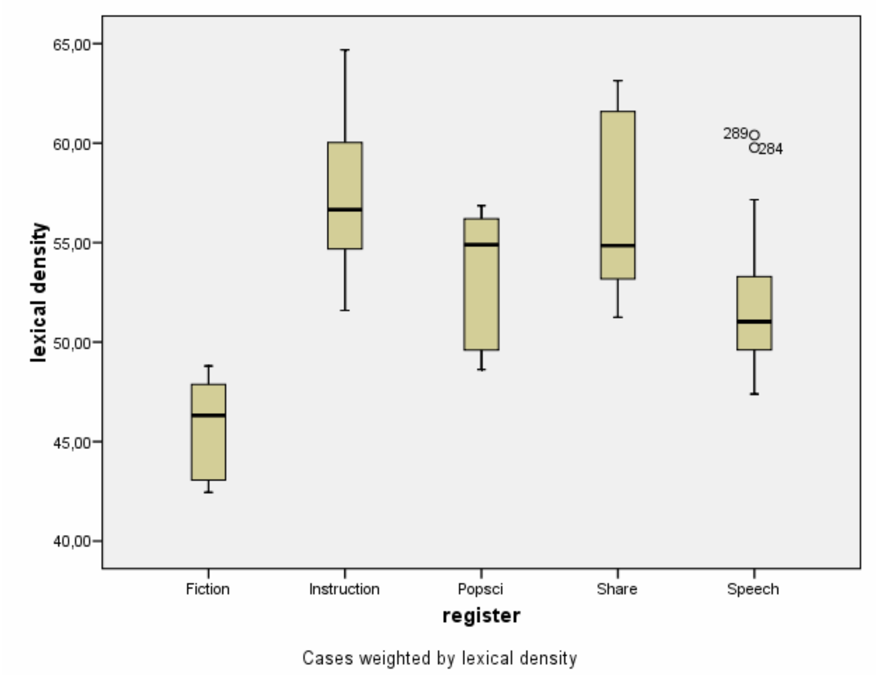


Figure 12: Variation of lexical density in five English original registers

One would expect the register of prepared speeches to exhibit a low lexical density as an indication of their spoken realization (this would be in line with Biber's, 1995 interpretation of the spoken texts in his corpus). While this register has the second lowest mean value (52.58 percent), it still has clearly higher values than FICTION. Figure 12 shows that English original fictional texts have the lowest maximum value and the lowest range. They also have the lowest mean value (45.72 percent) for lexical density. The minimum values for the four other registers displayed in the figure are either level with or higher than the maximum value for FICTION. The reason for these findings is probably that, on the one hand, the speeches carry more characteristics of the written mode due to their elaborated preparation process. On the other hand, the fictional texts frequently contain dialogical elements between the characters in the fictional world of the text. While these dialogues are invented and written by the author, they still contribute to spoken traits of the fictional texts.

The research queries presented in this paper give but a first impression of the range of the insights possible with the help of quantitative data containing rich linguistic annotation. Further queries comprising all other subdimensions of register analysis will result in a comprehensive overview of register variation based on theoretically-motivated criteria.

5. Conclusion

The research described here illustrates the use of linguistically annotated corpora across languages. In the application-oriented part, we have shown how corpora enriched with linguistic information facilitate solving typical translation problems in the language pair English-German, such as English cleft constructions. Information about part-of-speech, grammatical functions, phrase structure as well as alignment on different levels can help to identify the typical constructions. This type of translation memory look-up enables the translator to search for various problematic lexico-grammatical constructions and their aligned translation.

We also discussed findings available for linguistic research from these annotated corpora. The integration of these findings in applications for practical translators may prove a valuable resource for the practitioner.

The MySQL storage of linguistically annotated data combined with the possibility to exploit this data allows the easy and fast extraction of grammatically complex structures across languages and thus prepares the ground for their examination. Future work includes the development of a user-friendly interface that provides access to the queries discussed here for the ordinary user.

Acknowledgements

The work presented here is funded by the German Research Foundation (DFG) as project no. STE 840/5-1.

References

Baker, M. (1996) Corpus-based translation studies: The challenges that lie ahead, in H. Somers (ed.) Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager. Amsterdam: Benjamins, 175–86.

Brants, T. (2000) TnT - A Statistical Part-of-Speech Tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.

Biber, D. (1993) ‘Representativeness in Corpus Design’. *Literary and Linguistic Computing 8/4*, 243–57.

Biber, D. (1995) Dimensions of register variation. Cambridge: Cambridge Univ. Press.

Halliday, M.A.K. and R. Hasan (1989) Language, Context, and Text: Aspects of Language in a Social Semiotic Perspective. Oxford: Oxford Univ. Press.

Halliday, M.A.K. and J.R. Martin (1993) Writing Science: Literacy and Discursive Power. London, Washington, D.C.: Falmer Press.

Hansen, S. (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, S., S. Neumann and E. Steiner (to appear) ‘Cohesive explicitness and explicitation in an English-German translation corpus’. *Languages in Contrast 7:2 (2007).*

Hansen-Schirra, S., S. Neumann and M. Vela (2006) Multi-dimensional Annotation and Alignment in an English-German Translation Corpus. *Proceedings des Workshops Multi-dimensional Markup in Natural Language Processing (NLPXML-2006)* Trento, Italy, 4 April 2006, pp. 35–42.

Hawkins, J. (1986) A comparative typology of English and German. Unifying the contrasts. London: Croom Helm.

Heyn, M. (1996) Integrating machine translation into translation memory systems. *European Association for Machine Translation - Workshop Proceedings*, pp. 111–23. Geneva: ISSCO.

Maas, H. D. (1998) Multilinguale Textverarbeitung mit MPRO. *Europäische Kommunikationskybernetik heute und morgen '98*, Paderborn.

Sperberg-McQueen, C.M. and L. Burnard (eds.) (1994) Guidelines for Electronic Text Encoding and Interchange (TEI P3). Chicago and Oxford: Text Encoding Initiative.

Morris, J. and G. Hirst (1991) ‘Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text’. *Computational Linguistics 17:1*, 21–48.

Müller, C. and M. Strube (2003) Multi-Level Annotation in MMAX. *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pp. 198–207. Sapporo, Japan.

Neumann, S. and S. Hansen-Schirra (2005) The CroCo Project: Cross-linguistic corpora for the investigation of explicitation in translations. *Proceedings from the Corpus Linguistics Conference Series*, Vol. 1, no. 1, ISSN 1747–9398.

Och, F.-J. and H. Ney (2003) 'A Systematic Comparison of Various Statistical Alignment Models'. *Computational Linguistics 29:1*, 19–51.

Sampson, G. (1995) English for the Computer. The Susanne Corpus and Analytic Scheme. Oxford: Clarendon Press.

Santini, M. (2006) ‚Web pages, text types, and linguistic features: Some issues'. *ICAME Journal 30*, 67–86.

Schiller, A., S. Teufel and C. Stöckert (1999) Guidelines für das Tagging deutscher Textkorpora mit STTS, University of Stuttgart and Seminar für Sprachwissenschaft, University of Tübingen.

Schrader, B. (2006) ATLAS -- a new text alignment architecture. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, July 2006, Sydney, Australia, pp. 715-722. Association for Computational Linguistics.

Scott, M. (1996) WordSmith Tools Manual. Oxford: Oxford Univ. Press.

Steiner, E. (2004) Translated Texts: Properties, Variants, Evaluations. Frankfurt/M.: Peter Lang.

Steiner, E. (to appear) Empirical studies of translations as a mode of language contact - "explicitness" of lexicogrammatical encoding as a relevant dimension, in P. Siemund and N. Kintana (eds.), Language Contact and Contact Languages. Amsterdam: Benjamins.

Ure, J. (1971) Lexical density and register differentiation, in G.E. Perren and J.K. Trim (eds.) Applications of Linguistics, pp. 443-451. Cambridge: Cambridge Univ. Press.

Ventola, E. (1996) Packing and Unpacking of Information in Academic Texts, in E. Ventola and A. Mauranen (eds.) Academic Writing. Intercultural and Textual Issues, pp. 153–94. Amsterdam, Philadelphia: Benjamins.

Vinay, J.P. and J. Darbelnet (1995) A comparative stylistics of French and English. A methodology for translation. Amsterdam, Philadelphia: Benjamins.