Multi-dimensional Annotation and Alignment in an English-German Translation Corpus

Silvia Hansen-Schirra
Computational Linguistics & Applied Linguistics, Translation and Interpreting
Saarland University, Germany
hansen@coli.uni-sb.de

Stella Neumann
Applied Linguistics, Translation and Interpreting
Saarland University, Germany
st.neumann@mx.uni-saarland.de

Mihaela Vela
Applied Linguistics, Translation and Interpreting
Saarland University, Germany
m.vela@mx.uni-saarland.de

Abstract

This paper presents the compilation of the CroCo Corpus, an English-German translation corpus. Corpus design, annotation and alignment are described in detail. In order to guarantee the searchability and exchangeability of the corpus, XML stand-off mark-up is used as representation format for the multi-layer annotation. On this basis it is shown how the corpus can be queried using XQuery. Furthermore, the generalisation of results in terms of linguistic and translational research questions is briefly discussed.

1 Introduction

In translation studies the question of how translated texts differ systematically from original texts has been an issue for quite some time with a surge of research in the last ten or so years. Example-based contrastive analyses of small numbers of source texts and their translations had previously described characteristic features of the translated texts, without the availability of more large-scale empirical testing. Blum-Kulka (1986), for instance, formulates the hypothesis that explicitation is a characteristic phenomenon of translated versus original texts on the basis of linguistic evidence from individual sample texts showing that translators explicitate optional cohesive markers in the target text not realised in the source text. In general, explicitation covers all features that make implicit information in the source text clearer and thus explicit in the translation (cf. Steiner 2005).

Building on example-based work like Blum-Kulka's, Baker put forward the notion of translation universals (cf. Baker 1996) which can be analysed in corpora of translated texts regardless of the source language in comparison to original texts in the target language. Olohan and Baker (2000) therefore analyse explicitation in English translations concentrating on the frequency of the optional *that* versus zero-connector in combination with the two verbs *say* and *tell*. While being extensive enough for statistical interpretation, corpus-driven research like Olohan and Baker's is limited in its validity to the selected strings.

More generally speaking, there is a gap between the abstract research object and the low level features used as indicators. This gap can be reduced by operationalising notions like explicitation into syntactic and semantic categories, which can be annotated and aligned in a corpus. Intelligent queries then produce linguistic evidence with more explanatory power than low level data obtained from raw corpora. The results are not restricted to the queried strings but extend to more complex units sharing the syntactic and/or semantic properties obtained by querying the annotation.

This methodology serves as a basis for the CroCo project, in which the assumed translation property of explicitation is investigated for the language pair English – German. The empirical evidence for the investigation consists in a corpus 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). Alto-

gether the CroCo Corpus comprises one million words. Additionally, reference corpora are included for German and English. The reference corpora are register-neutral including 2,000 word samples from 17 registers (see Neumann & Hansen-Schirra 2005 for more details on the CroCo corpus design).

The CroCo Corpus is tokenised and annotated for part-of-speech, morphology, phrasal categories and grammatical functions. Furthermore, the following (annotation) units are aligned: words, grammatical functions, clauses and sentences. The annotation and alignment steps are described in section 2.

Each annotation and alignment layer is stored separately in a multi-layer stand-off XML representation format. In order to empirically investigate the parallel corpus (e.g. to find evidence for explicitation in translations), XQuery is used for posing linguistic queries. The query process itself works on each layer separately, but can also be applied across different annotation and alignment layers. It is described in more detail in section 3. This way, parallel text segments and/or parallel annotation units can be extracted and compared for translations and originals in German and English.

2 CroCo XML

The annotation in CroCo extends to different levels in order to cover possible linguistic evidence on each level. Thus, each kind of annotation (part-of-speech, morphology, phrase structure, grammatical functions) is realised in a separate layer. An additional layer is included which contains comprehensive metainformation in separate header files for each text in the corpus. The file containing the indexed tokens (see section 2.1) includes an **xlink** attribute referring to this header file as depicted in Figure 2.1. The metadata are based on the TEI guidelines[1] and include register information. The complex multilingual structure of the corpus in combination with the multi-layer annotation requires indexing the corpus. The indexing is carried out on the basis of the tokenised corpus. Index and annotation layers are kept separate using XML stand-off mark-up. The mark-up builds on XCES[2]. Different formats of the multiple annotation and alignment outputs are converted with Perl scripts. Each annotation and alignment unit is indexed. The respective annotations and alignments are linked to the indexed units via XPointers.

The following sections describe the different annotation layers and are exemplified for the German original sentence in (1) and its English translation in (2)[3].

(1) Ich spielte viele Möglichkeiten durch, stellte mir den Täter in verschiedenen Posen vor, ich und die Pistole, ich und die Giftflasche, ich und der Knüppel, ich und das Messer.

(2) I ran through numerous possibilities, pictured the perpetrator in various poses, me with the gun, me with the bottle of poison, me with the bludgeon, me with the knife.

2.1 Tokenisation and indexing

The first layer to be presented here is the tokenisation layer. Tokenisation is performed in CroCo for both German and English by TnT (Brants 2000), a statistical part-of-speech tagger. As shown in Figure 2.1 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 file is an original or a translation.

```
<document
xmlns:xlink="http://www.w3.org/1999/
xlink" name="GO.tok.xml" xml:lang="de"
docType="ori">
<header xlink:href="GO.header.xml"/>
 <tokens>
  <token id="t64" strg="Ich"/>
  <token id="t65" strg="spielte"/>
  <token id="t66" strg="viele"/>
  <token id="t67"
   strg="Möglichkeiten"/>
  <token id="t68" strg="durch"/>
  <token id="t69" strg=","/>
</tokens>
</document>
```

Figure 2.1. Tokenisation and indexing

Similar index files necessary for the alignment of the respective levels are created for the units chunk, clause and sentence. These units stand in

[1] http://www.tei-c.org
[2] http://www.xml-ces.org

[3] All examples are taken from the CroCo Corpus.

a hierarchical relation with sentences consisting of clauses, clauses consisting of chunks etc.

2.2 Part-of-speech tagging

The second layer annotated for both languages is the part-of-speech layer, which is provided again by TnT[4]. 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 2.2. 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. For example, the German token pointing to "t65" in the token index file whose **strg** value is *stellte* is a finite verb (with the PoS tag vvfin).

```
<document
xmlns:xlink="http://www.w3.org/1999/
xlink" name="GO.tag.xml">
 <tokens xml:base="GO.tok.xml">
  <token pos="pper" xlink:href="#t64"/>
  <token pos="vvfin"
   xlink:href="#t65"/>
  <token pos="pidat"
   xlink:href="#t66"/>
  <token pos="nn" xlink:href="#t67"/>
  <token pos="ptkvz"
   xlink:href="#68"/>
  <token pos="yc" xlink:href="#t69"/>
</tokens>
</document>
```

Figure 2.2. PoS tagging

2.3 Morphological annotation

Morphological information is particularly relevant for German due to the fact that this language carries much syntactic information within morphemes rather than in separate function words like English. Morphology is annotated in CroCo with MPro, a rule-based morphology tool (Maas 1998). This tool works on both languages. As shown in Figure 2.3 each token has morphological attributes such as **person**, **case**, **gender**, **number** and **lemma.** As before, the **xlink** attribute refers back to the index file, thus providing the connection between the morphological attributes and the **strg** information in the index file.

For the morphological annotation of the German token "t65" in Figure 2.3 the **strg** value is determined by following the XPointer "t65" to the token index file, i.e. *spielte*. The **pos** value is retrieved by searching in the tag annotation for the file with the same **xml:base** value. The matching tag, in this case vvfin, carries the same XPointer “t65”.

```
<document
xmlns:xlink="http://www.w3.org/1999/
xlink" name="GO.morph.xml">
 <tokens xml:base="GO.tok.xml">
  <token strg="Ich" per="1" case="nom"
   nb="sg" gender="f;m" lemma="ich"
   lb="ich" xlink:href="#t64"/>
  <token strg="spielte" vtype="fiv"
   tns="past" per="3" nb="sg"
   lemma="spielen" lb="spielen" comp=
   "spielen" xlink:href="#t65"/>
  <token strg="viele" case="nom;acc"
   nb="plu" gender="f" lemma="viel"
   lb="viel" comp="viel" deg="base"
   xlink:href="#t66"/>
  <token strg="Möglichkeiten" case=
   "nom;acc" nb="plu" gender="f" lemma=
   "möglichkeit" lb="möglich" comp=
   "möglichkeit" xlink:href="#t67"/>
  <token strg="durch" lemma="durch"
   lb="durch" pref="vzs"
   xlink:href="#t68"/>
  <token strg="," lemma="," lb=","
   xlink:href="#t69"/>
 </tokens>
</document>
```

Figure 2.3. Morphological annotation

2.4 Phrase chunking and annotation of grammatical functions

Moving up from the token unit to the chunk unit, first we have to index these units again before we can annotate them. The chunk index file assigns an **id** attribute to each chunk within the file. The problem of discontinuous phrase chunks is solved by listing child tags referring to the individual tokens which make up the chunk via **xlink** attributes. Figure 2.4 shows that the VP “ch14” in the German phrase annotation consists of “t70” (*stellte*) and “t77” (*vor*).

```
<document xmlns:xlink=
"http://www.w3.org/1999/xlink"
name="GO.chunk.xml">
 <chunks xml:base="GO.tok.xml">
  <chunk id="ch13">
   <tok xlink:href="#t66"/>
   <tok xlink:href="#t67"/>
  </chunk>
  <chunk id="ch14">
   <tok xlink:href="#t70"/>
  </chunk>
  <chunk id="ch15">
   <tok xlink:href="#t71"/>
  </chunk>
  <chunk id="ch16">
   <tok xlink:href="#t72"/>
   <tok xlink:href="#t73"/>
  </chunk>
```

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

```
  <chunk id="ch17">
   <tok xlink:href="#t74"/>
   <chunk id="ch18">
    <tok xlink:href="#t75"/>
    <tok xlink:href="#t76"/>
   </chunk>
  </chunk>
  <chunk id="ch19">
    <tok xlink:href="#t77"/>
  </chunk>
 </chunks>
</document>
```

Figure 2.4. Chunk indexing

The phrase structure annotation (see Figure 2.5) assigns the **ps** 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, as our focus within this layer is more on complete phrase chunks and their grammatical functions.

```
<document
 xmlns:xlink="http://www.w3.org/1999/
 xlink" name="GO.ps.xml">
 <chunks xml:base="GO.chunk.xml">
  <chunk ps="NP" xlink:href="#ch13"/>
  <chunk ps="VPFIN"
   xlink:href="#ch14"/>
  <chunk ps="NP" xlink:href="#ch15"/>
  <chunk ps="NP" xlink:href="#ch16"/>
  <chunk ps="PP" xlink:href="#ch17"/>
  <chunk ps="NP" xlink:href="#ch18"/>
  <chunk ps="VPPRED"
   xlink:href="#ch19"/>
 </chunks>
</document>
```

Figure 2.5. Phrase structure annotation

The annotation of grammatical functions is again kept in a separate file (see Figure 2.6). Only the highest phrase nodes are annotated for their grammatical function with the attribute **gf**. The XPointer links the annotation of each function to the chunk **id** in the chunk index file. From this file in turn the string can be retrieved in the token annotation. For example, the English chunk "ch13" carries the grammatical function of direct object (DOBJ). It is identified as an NP in the phrase structure annotation by comparing the **xml:base** attribute value of the two files and the XPointers.

```
<document
xmlns:xlink="http://www.w3.org/1999/
xlink" name="GO.gf.xml">
 <chunks xml:base="GO.chunk.xml">
  <chunk gf="DOBJ" xlink:href="#ch13"/>
  <chunk gf="FIN" link:href="#ch14"/>
  <chunk gf="IOBJ" xlink:href="#ch15"/>
  <chunk gf="DOBJ" xlink:href="#ch16"/>
  <chunk gf="ADV" xlink:href="#ch17"/>
  <chunk gf="PRED" xlink:href="#ch19"/>
 </chunks>
</document>
```

Figure 2.6. Annotation of grammatical functions

2.5 Alignment

In the examples shown so far, the different annotation layers linked to each other all belonged to the same language. By aligning words, grammatical functions, clauses and sentences, the connection between original and translated text is made visible. The use of this multi-layer alignment will become clearer from the discussion of a sample query in section 3.

For the purpose of the CroCo project word alignment is realised with GIZA++ (Och & Ney 2003), a statistical alignment tool. Chunks and clauses are aligned manually with the help of MMAX II (Müller & Strube 2003), a tool allowing assignment of own categories and linking units. Finally, sentences are aligned using WinAlign, an alignment tool within the Translator's Workbench by Trados (Heyn 1996).

The alignment procedure produces four new layers. It follows the XCES standard. Figure 2.7 shows the chunk alignment of (1) and (2). In this layer, we align on the basis of grammatical functions instead of phrases since this annotation includes the information of the phrase chunking as well as on the semantic relations of the chunks. The grammatical functions are mapped onto each other cross-linguistically and then aligned according to our annotation and alignment scheme. The **trans.loc** attribute locates the chunk index file for the aligned texts in turn. 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 chunk pointing to the chunk index file. As can be seen from Figure 2.7, chunks which do not have a matching equivalent receive the value "#undefined", a phenomenon that will be of interest in the linguistic interpretation on the basis of querying the corpus.

```
<document
xmlns:xlink="http://www.w3.org/1999/
xlink" name="gfAlign.xml">
<translations xml:base="/CORPUS/">
<translation trans.loc="GO.chunk.xml"
xml:lang="de" n="1"/>
<translation
trans.loc="ETrans.chunk.xml"
xml:lang="en" n="2"/>
```

```
</translations>
 <chunks>
  <chunk>
   <align xlink:href="#ch14"/>
   <align xlink:href="#ch16"/>
  </chunk>
  <chunk>
   <align xlink:href="#ch15"/>
   <align xlink:href="#undefined"/>
  </chunk>
  <chunk>
   <align xlink:href="#ch16"/>
   <align xlink:href="#ch17"/>
  </chunk>
  <chunk>
   <align xlink:href="#ch17"/>
   <align xlink:href="#ch18"/>
  </chunk>
  <chunk>
   <align xlink:href="#ch19"/>
   <align xlink:href="#undefined"/>
  </chunk>
 </chunks>
</document>
```

Figure 2.7. Chunk alignment

3 Querying the CroCo Corpus

The comprehensive annotation including the alignment described in section 2 is the basis for the interpretation to be presented in what follows. We concentrate on two types of queries into the different alignment layers that are assumed relevant in connection with our research question.

3.1 Crossing lines and empty links

From the linguistic point of view we are interested in those units in the target text which do not have matches in the source text and vice versa, i.e. **empty links**, or whose alignment crosses the alignment of a higher level, i.e. **crossing lines**. We analyse for instance stretches of text contained in one sentence in the source text but spread over two sentences in the target text, as this probably has implications for the overall information contained in the target text. We would thus pose a query retrieving all instances where the alignment of the lower level is not parallel to the higher level alignment but points into another higher level unit. In the example below the German source sequence (3) as well as the English target sequence (4) both consist of three sentences. These sentences are each aligned as illustrated by dashed boxes in Figure 3.1.

```
(3) Aus dem Augenwinkel sah ich,
wie eine Schwester dem Bettnachbarn
das Nachthemd wechselte. Sie rieb
den Rücken mit Franzbranntwein ein
und massierte den etwas jüngeren
Mann, dessen Adern am ganzen Körper
bläulich hervortraten. Ihre Hände
ließen ihn leise wimmern.

(4) Out of the corner of my eye I
watched a nurse change his
neighbor's nightshirt and rub his
back with alcoholic liniment. She
massaged the slightly younger man,
whose veins stood out blue all over
his body. He whimpered softly under
her hands.
```

In German the first two sentences are subdivided into two clauses each. The English target sentences are co-extensive with the clauses contained in each sentence. This means that two English clauses have to accommodate four German clauses. Figure 3.1 shows that the German clause 3 (*Sie rieb den Rücken mit Franzbranntwein ein*) in sentence 2 is part of the bare infinitive complementation (*…and rub his back with alcoholic liniment*) in the English sentence 1. The alignment of this clause points out of the aligned first sentence, thus constituting crossing lines.

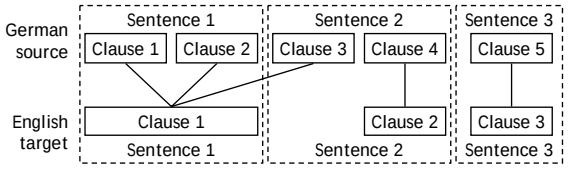

Figure 3.1. Sentence and clause alignment

The third sentence also contains a crossing line, in this case on the levels of chunk and word alignment: The words *Ihre Hände* in the German subject chunk are aligned with the words *her hands* in the English adverbial chunk. However, this sentence is particularly interesting in view of empty links. The query asks for units not matching any unit in the parallel text, i.e. for **xlink** attributes whose values are "#undefined" (cf. section 2.5). In Figure 3.2, the empty links are marked by a black dot.

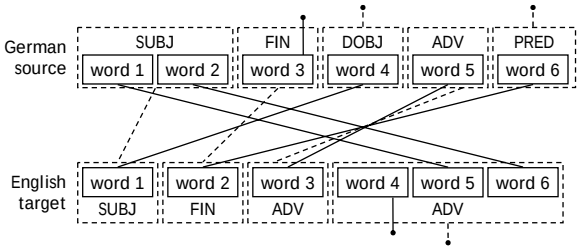

Figure 3.2. Chunk and word alignment

Our linguistic interpretation is based on a functional view of language. Hence, the finite *ließen* (word 3) in the German sentence is interpreted as a semi-auxiliary and thus as the finite part of the verbal group. Therefore, *wimmern* (word 6) receives the label PRED, i.e. the non-finite part of the verb phrase, in the functional analysis. This German word is linked to word 2 (*whimpered*) in the target sentence, which is assigned FIN, i.e. the finite verb in the layer of grammatical functions. As FIN exists both in the source and in the target sentences, this chunk is aligned. The German functional unit PRED does not have an equivalent in the target text and gets an empty link. Consequently, word 3 in the source sentence (*ließen*) receives an empty link as well. This mismatch will be interpreted in view of our translation-oriented research. In the following subsection we will see how these two phenomena can be retrieved automatically.

3.2 Corpus exploitation using XQuery

Since the multi-dimensional annotation and alignment is realised in XML, the queries are posed using XQuery[5]. This query language is particularly suited to retrieve information from different sources like for instance individual annotation and alignment files. The use for multi-layer annotation is shown in (Teich et al. 2001).

The query for determining an empty link at word level can be formulated as follows: find all words which do not have an aligned correspondent, i.e. which carry the **xlink** attribute value "#undefined". The same query can be applied on the chunk level, the query returning the grammatical functions that do not have an equivalent in the other language.

(5)Ihre Hände ließen ihn leise wimmern.

(6) He whimpered softly under her hands.

Applied to the sentences in (5) and (6) the XQuery in Figure 3.3 returns all German and English words, which receive an empty link due to a missing equivalent in alignment (*ließen* and *under*). This query can be used analogously in all other alignment layers. It implies the call of a self-defined XQuery function (see Figure 3.4), which looks in the correspondent index file for words not aligned.

[5] http://www.w3.org/TR/xquery

```
let $doc := .
for $k in $doc//tokens/token
return
 if ($k/align[1][@xlink:href="#un-
  defined"] and $k/align[2]
  [@xlink:href!="#undefined"])
 then local:getString($k/align[1]/
  @xlink:href,$k/align[2]/@xlink:href,
  $doc//translations/translation
  [@n='2']/@trans.loc)
 else if ($k/align[1][@xlink:href
  !="#undefined"] and $k/align[2]
  [@xlink:href="#undefined"])
 then local:getString($k/align[1]/
  @xlink:href,$k/align[2]/@xlink:href,
  $doc//translations/translation
  [@n='1']/@trans.loc)
 else ()
```

Figure 3.3. XQuery for empty links

```
declare function local:getString
($firstToken as xs:string,$secondTo-
ken as xs:string,$fileName as
xs:string) as element()
 {let $res:=(if(($firstToken eq
  "#undefined") and ($lang eq doc
  ($fileName)//document/@xml:lang))
  then doc($fileName)//tokens/token[@id
  eq substring-after($secondToken,"#")]
  else if (($secondToken eq "#unde-
  fined") and ($lang eq doc($fileName)
  //document/@xml:lang))
  then doc($fileName)//tokens/token[@id
  eq substring-after($firstToken,"#")]
  else ())
return
 <token>{$res/@strg}</token>};
```

Figure 3.4. XQuery function for missing alignment

Querying crossing lines in the German source sentence in (5) and the English target sentence in (6) is based on the annotation at word level as well as on the annotation at the chunk level. As mentioned in section 3.1, crossing lines are identified in (5) and (6) if the words contained in the chunks aligned on the grammatical function layer are not aligned on the word level. This means that the German subject is aligned with the English subject, but the words within the subject chunk are aligned with words in other grammatical functions instead.

In a first step, the query for determining a crossing line requires information about all aligned German chunks with a **xlink** attribute whose value is not "#undefined" and all aligned German words with a **xlink** attribute whose value is not "#undefined". Then all German words that are not aligned on the word level but are aligned as part of chunks on the chunk level

are filtered out. Figure 3.6 reflects the respective XQuery.

```
let $doc := .
for $k in $doc//chunks/chunk
let $ch1:=(if($k/align[1][@xlink:href
 !="#undefined"] and $k/align[2]
 [@xlink:href!="#undefined"])
 then doc($doc//translations/trans-
 lation[@n='1']/@trans.loc)//chunks
 /chunk[@id eq substring-after
 ($k/align[1]/@xlink:href,"#")]
 else ())
let $ch2:=(if($k/align[1][@xlink:href
 !="#undefined"] and $k/align[2]
 [@xlink:href!="#undefined"])
 then (doc($doc//translations/transla-
 tion[@n='2']/@trans.loc)//chunks/chunk
 [@id eq substring-after($k/align[2]/
 @xlink:href,"#")])
 else ())
for $i in doc("g2e.tokenAlign.xml")
 //tokens/token
let $tok1:=(if($i/align[1][@xlink:href
 !="#undefined"] and $i/align[2]
 [@xlink:href!="#undefined"])
 then(doc(doc("g2e.tokenAlign.xml")
 //translations/translation[@n='1']
 /@trans.loc)//tokens/token[@id eq
 substring-after($i/align[1]
 /@xlink:href,"#")])
 else ())
let $tok2:=(if($i/align[1][@xlink:href
 !="#undefined"] and $i/align[2]
 [@xlink:href!="#undefined"])
 then(doc(doc("g2e.tokenAlign.xml")
 //translations/translation[@n='2']
 /@trans.loc)//tokens/token[@id eq
 substring-after($i/align[2]
 /@xlink:href,"#")])
 else ())
where(local:containsToken($ch1/tok
 [position()=1],$ch1/tok[last()],$tok1
 /@id) and not(local:containsToken
 ($ch2/tok[position()=1],
 $ch2/tok[last()],$tok2/@id)))
return $tok1
```

Figure 3.6. XQuery for crossing lines

First, the aligned chunks (**$ch1** and **$ch2**) are saved into variables. These values are important in order to detect the span for each of the chunks (**$ch1/tok[position()=1], $ch1/tok[last()]** and **$ch2/tok[position()=1], $ch2/tok[last()]**), and to identify the words making up the source chunks as well as their German or English equivalents. In the second step all words that do not have empty links are saved (**$tok1** and **$tok2**). The last step filters the crossing lines, i.e. word alignments pointing out of the chunk alignment. For this purpose, we define a new function (**local:containsToken**) which tests whether a word belongs to a chunk or not. By applying **local:containsToken** for the German original and **not(local:containsToken)** for the English translation, all words in the German chunks whose aligned English equivalent words do not belong to the aligned English chunks are retrieved. The example query returns the German words *Ihre Hände* that are part of the German subject chunk and which are aligned with the English words *her hands* that again are part of the second adverbial chunk.

4 Summary and conclusions

In a broader view, it can be observed that there is an increasing need in richly annotated corpora across all branches of linguistics. The same holds for linguistically interpreted parallel corpora in translation studies. Usually, though, the problem with large-scale corpora is that they do not reflect the complexity of linguistic knowledge we are used to dealing with in linguistic theory. Simple research questions can of course be answered on the basis of raw corpora or with the help of an automatic part-of-speech tagging. Most linguistics and translation scholars are, however, interested in more complex questions like the interaction of syntax and semantics across languages.

The research described here shows the use of comprehensive multi-layer annotation across languages. By relating a highly abstract research question to multiple layers of lexical and grammatical realisations, characteristic patterns of groups of texts, e.g. explicitation in translations and originals in the case of the CroCo project, can be identified on the basis of statistically relevant linguistic evidence.

If we want to enrich corpora with multiple kinds of linguistic information, we need a linguistically motivated model of the linguistic units and relations we would like to extract and draw conclusions based on an annotated and aligned corpus. So the first step for the compilation of a parallel translation corpus is to provide a classification of linguistic units and relations and their mappings across source and target languages. The classification of English and German linguistic units and relations chosen for the CroCo project (i.e. for the investigation of explicitation in translations and originals) is reflected in the CroCo annotation and alignment schemes and thus in the CroCo Corpus annotation and alignment.

From a technical point of view, the representation of a multilingual resource comprehensively

annotated and aligned is to be realised in such a way that

- multiple linguistic perspectives on the corpus are possible since different annotations and alignments can be investigated independently or in combination,
- the corpus format guarantees best possible accessibility and exchangeability, and
- the exploitation of the corpus is possible using easily available tools for search and analysis.

We coped with this challenge by introducing a multi-layer stand-off corpus representation format in XML (see section 2), which takes into account not only the different annotation layers needed from a linguistic point of view, but also multiple alignment layers necessary to investigate different translation relations.

We also showed how the CroCo resource can be applied to complex research questions in linguistics and translation studies using XQuery to retrieve multi-dimensional linguistic information (see section 3). Based on the stand-off storage of annotation and alignment layers combined with the possibility to exploit the required layers through intelligent queries, parallel text segments and/or parallel annotation units can be extracted and compared across languages.

In order to make the CroCo resource available to researchers not familiar with the complexities of XML mark-up and the XQuery language, a graphical user interface will be implemented in Java which allows formulating queries without knowledge of the XQuery syntax.

Acknowledgement

The authors would like to thank the reviewers for their excellent comments and helpful feedback on previous versions of this paper.

The research described here is sponsored by the German Research Foundation as project no. STE 840/5-1.

References

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers (ed.). *Terminology, LSP and Translation.* Benjamins, Amsterdam:175-186.

Douglas Biber. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8/4:243-257.

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in Translation. In Juliane House and Shoshana Blum-Kulka (eds.). *Interlingual and Intercultural Communication.* Gunter Narr, Tübingen:17-35.

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

Matthias Heyn. 1996. Integrating machine translation into translation memory systems. *European Association for Machine Translation - Workshop Proceedings*, ISSCO, Geneva:111-123.

Heinz Dieter Maas. 1998. Multilinguale Textverarbeitung mit MPRO. *Europäische Kommunikationskybernetik heute und morgen '98*, Paderborn.

Christoph Müller and Michael Strube. 2003. Multi-Level Annotation in MMAX. *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan:198-107.

Stella Neumann 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.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Journal of Computational Linguistics* Nr.1, vol. 29:19-51.

Maeve Olohan and Mona Baker. 2000. Reporting *that* in Translated English. Evidence for Subconscious Processes of Explicitation? *Across Languages and Cultures* 1(2):141-158.

Geoffrey Sampson. 1995. *English for the Computer. The Susanne Corpus and Analytic Scheme*. Clarendon Press, Oxford.

Anne Schiller, Simone Teufel and Christine Stöckert. 1999. Guidelines für das Tagging deutscher Textkorpora mit STTS, *University of Stuttgart and Seminar für Sprachwissenschaft*, University of Tübingen.

Erich Steiner. 2005. Explicitation, its lexicogrammatical realization, and its determining (independent) variables – towards an empirical and corpus-based methodology. *SPRIKreports* 36:1-43.

Elke Teich, Silvia Hansen, and Peter Fankhauser. 2001. Representing and querying multi-layer annotated corpora. *Proceedings of the IRCS Workshop on Linguistic Databases*. Philadelphia: 228-237.