The CroCo Project

Cross-linguistic corpora for the investigation of explicitation in translations

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

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

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 (cf. for instance Blum-Kulka 1986). Building on these studies, Mona 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. Among the proposed properties are the following: translations are, broadly speaking, simpler than originals ("simplification"), information is spelt out explicitly in the translated text ("explicitation") and translators replace untypical features by typical features of the target language ("normalisation"; "law of growing standardisation", cf. Toury 1995). Finally, when comparing a corpus of translations with a corpus of originals in the same language, the linguistic features found in the translation corpus are expected to concentrate around the most frequently used options while the corpus of originals displays more variation ("levelling out"). Beyond these properties, which can all be detected by comparing translations with originals in the same language, Gideon Toury describes the influence of the source language in the translation ("law of interference", Toury 1995; Teich 2003 offers a corpus-based analysis of what she calls "shining through").

The CroCo project picks out the assumed property of explicitation and will investigate it for the language pair English – German in the light of three factors which we consider relevant for an in-depth investigation of this translation property. The existing studies dealing with explicitation in translation suggest that indicators can be found on each linguistic level. Furthermore, Steiner (2001a) points out that, in order to elucidate the phenomenon, it is necessary to take into account at least three sources of explanation: language typology, the contrastive features of the (activated) register and the translation process. Finally, previous studies like Teich (2003) have shown that not all of the properties are universal but that they may depend on the translation direction, the language pair or otherwise have a limited scope. Figure 1 shows a matrix combining all three factors which have to be taken into account for a comprehensive examination of explicitation.

* We would like to thank Erich Steiner, Kerstin Kunz, Andrea Kamm and Elke Teich for fruitful discussions. The research described here is supported by the DFG grant no. STE 840/5-1.

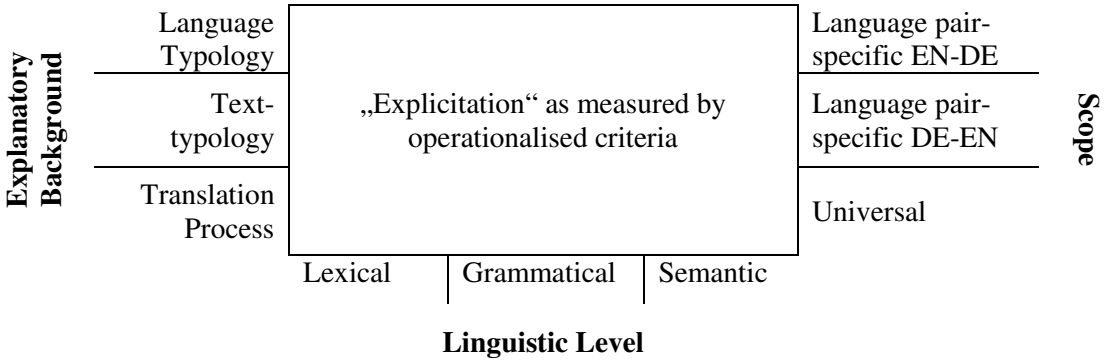

Figure 1. Matrix combining all factors relevant for interpreting explicitation

The remainder of the paper is organised as follows. First, we will explain the property of explicitation in more detail (Section 2). Then we will move on to present the corpus design needed to analyse explicitation in the light of the three factors mentioned above (Section 3). In Section 4, the annotation and mark-up of our corpus is described. And finally, in Section 5, we will widen the perspective to future extensions on the basis of the corpus thus built and annotated.

2 Why more on explicitation in translations

As early as 1958 Vinay & Darbelnet mentioned explicitation as a stylistic translation technique (taken from the English translation from 1995). This represents a part of translation studies which focuses on individual translation strategies in an example-based way. Blum-Kulka's explicitation hypothesis (1986) was the starting point for empirical research into this property of translations, defining explicitation as a characteristic phenomenon of translated versus original texts. Blum-Kulka discusses the use of explicit cohesion markers in target texts resulting in increased text length. Baker takes up this feature of increased text length as an indicator for explicitation. It is her merit to have introduced corpus-linguistic methods to the study of translation properties (cf. Baker 1993; for an in-depth analysis of text length in connection with explicitation cf. Frankenberg-Garcia 2004).

The corpus queries Baker proposes operate on raw text corpora. This kind of corpora may be useful for developing assumptions about linguistic phenomena, but it limits the information that can be obtained to queries based on word strings. Olohan and Baker (2000) show how far you can go with string-based queries, but it also becomes obvious that the queries are restricted to a narrow set of strings, in their case the combination of the two verbs *say* and *tell* with the optional *that* versus zero-connector. It can be assumed that there are more meaningful indicators on the level of grammar, e.g. by abstracting from given lexical strings to categories of verbs. If we want to obtain more meaningful findings about these deeper structures of translations we first need to enrich the corpora with linguistic information. In this way, we are able to reduce the gap between highly abstract concepts like "explicitation" and very low level phenomena like text length in terms of word-forms or even characters, which may be subject to other influences.

A linguistically informed analysis of this kind combines with the linguistically richer approach to explicitation we propose. While we acknowledge the frequent use of additions to the target text, this phenomenon is not in the focus of our analysis since, strictly speaking, it does not belong to the scope of explicitation. Explicitation inherently means that the explicitated element must have been somehow implicitly present, i.e. linguistically traceable, in the source text (or, more generally speaking, in the implicit version). This has implications

for selecting features identified as representing explicitation: We only annotate those features whose linguistic realisation allows tracing implicit information in the source version. Thus example 1[1] below will not be counted, because the addition "der Stadt Calgary" (*of the city of Calgary)* cannot be attributed to linguistic features of the source version, but rather to some extra-linguistic knowledge available from the context. In example 2, on the other hand, the added information in the target version ("him") stems from the fact that the nominalisation in the source text includes an implicit semantic role which has to be spelt out in the finite construction of the target version. Furthermore, the finite verb in the target text contains tense information implicit in the source text. Finally, "Männer" (*men*) is specified as "father and son" in the target text. As can be seen from example 2, explicitating goes hand in hand with disambiguating the source text. While the source version allows at least two interpretations as to who had been called by the men, the translator has taken over the interpretation task and thus narrowed down the possible interpretations of the text, if we assume the extremely likely co-referentiality between "he" and "him".

English ST [on the Canadian city of Calgary]: *The tower also contains the city's tourist office.*
Modified German TT: *Im Turm ist auch das Tourismusbüro der Stadt Calgary untergebracht.*
English gloss: *In the tower is also the tourism office of the city of Calgary accommodated.*
Example 1

German ST: *Er erwähnte die Anrufe der beiden Männer...*
English gloss: *He mentioned the calls of the two men...*
English TT: *He mentioned that both, father and son, had called him...*
Example 2

In order to get a better understanding of the linguistic processes resulting in phenomena like this, we need a deeper analysis of possible indicators for explicitation on all linguistic levels. If we want to go beyond describing the phenomenon, we have to interpret the resulting data with respect to the abstract sources of explanation mentioned in Section 1. This in turn means that we have to build a complex corpus design allowing these explanations.

3 Corpus design

We have set ourselves the task of building a resource which has a representative size, which is well-balanced and which guarantees comparability across languages, targeting an overall size of 1 million words. On the content side, the CroCo project endeavours to cover features related to explicitation on all linguistic levels. Finally, the need to explain why explicitation takes place adds another dimension to the criteria for the corpus design. In this section we describe the resulting decisions for the CroCo Corpus.

First, with respect to corpus size, we face the problem that we cannot cover all texts in one corpus. Therefore we have to take a representative sample from the basic population of all texts. However, representativeness can only be achieved if the basic population can be determined. While we can, for instance, count all people living on a given stretch of earth, we cannot count all texts produced within a given period of time (if we do not want to narrow the sample down to a restricted author or author's collective). One might think, merely increasing the size of the resource as much as possible both in terms of text types covered and of number of words contained may ultimately equal representativeness. We content that a smaller corpus which is well-designed and annotated is preferable to a large one which may contain material not adding any information to the research question. And, of course, we do not have the

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

means to annotate a giga-word corpus with the linguistic information relevant for understanding explicitation.

Douglas Biber (1990, 1993) argues that smaller corpora – if well-balanced – are capable of covering all linguistic features of a given register. It may be helpful to approximate representativeness by making meaningful design decisions. In our case, this means choosing those registers significant for translation and drawing enough samples within one register in order to cover all relevant linguistic features. Biber's calculations, i.e. 10 texts per register with a length of 1,000 words (cf. Biber 1990, 1993), serve as an orientation for the size of our corpus. As to balance of the corpus, four criteria should be considered: publication date of the corpus candidates, regional language variety (not only English can be subdivided into a range of varieties but also German has at least three varieties), functional variety (register) and text length. Provided that the reference corpora, which constitute the basis of comparison for each language (see below), and the register-controlled corpora cover the same period of time, publication date should not be a decisive factor for the analysis of translation properties. In order to exclude any influence from this factor we take 1991 (the publication date of the FLOB-Corpus texts) as a starting date.

If research on translation properties is the main interest for building a corpus, balance with respect to language variety is not a hard criterion for the corpus design. Conversely, comparability across languages is an important and not trivial issue, particularly, if we aim at analysing register specificities as one factor for translation properties. Even in the language pair English and German, which is in close contact and where the languages are similarly specialised in terms of registers, there are potentially numerous registers which are not entirely comparable. In CroCo, the question of functional variety is therefore addressed in two steps. First, the decision which registers are included is based on registerial considerations (for a basic description of this kind of register analysis see Halliday & Hasan 1989): Each register should ideally vary from the other registers in one sub-dimension of the three register variables field, tenor and mode of discourse (cf. Steiner 2001b who deals with contrastive register analysis). Since such a fine-grained register classification is not available, we decided that each sub-dimension relevant in the context of written translation should be foregrounded in at least one register included in the corpus. This resulted in the decision to include 8 registers listed in Table 1.

Register	**Foregrounded sub-dimension**
popular-scientific texts	social role, experiential domain
tourism leaflets	goal orientation, experiential domain
prepared speeches	appraisal, medium, experiential domain
political essays on economics	appraisal, experiential domain
fictional texts	language role, experiential domain
corporate communication	social role, experiential domain
instruction manuals	goal orientation, language role, exp. domain
websites	social distance, channel, experiential domain

Table 1. Register variation in the CroCo Corpus

Two additional registers are included the text archive but are not processed in the first place because they are only available in one translation direction: court decisions (DE-EN) and scientific abstracts from the medical domain (EN-DE). In the second step, the intra- and interlingual comparability of the texts collected in each register is considered in the form of a

modest register analysis. The resulting register information is included in the metadata of the corpus. This allows us to filter the corpus according to specific register features.

Undoubtedly, it is desirable to collect full texts. However, features representing candidates for explicitation indicators on a deeper linguistic level typically can only be discovered on the basis of costly manual annotation. These indicators are the features the CroCo project is mainly interested in (see Section 4.3). Furthermore, the interpretation should not be limited to certain registers for no other reason than these registers consisting of short texts. Therefore, the CroCo Corpus is conceived as a dynamic resource which allows easy drawing of subcorpora comprising samples from longer texts for the purpose of small-scale manual analyses. Needless to say that text length may not be the only criterion for a subcorpus taken from the CroCo Corpus.

The last dimension influencing the corpus design is the explanatory background for the interpretation. If we want to trace back the reasons why we find translation properties like explicitation in translations we have to build a corpus which at least allows

- identifying so-called obligatory explicitation, i.e. those changes caused by differences in the language systems involved. These can only be retrieved by including both source and target language.
- comparing contrastive registers and thus distinguishing features which are due to specific register characteristics in the respective language.
- assigning the remaining cases of explicitation to the translation process proper by way of ruling out the other two factors.

We include reference corpora both in English (ER) and German (GR) for detecting contrastive restrictions of the respective language systems which force the translator to explicitate a source language structure. The reference corpora also allow identifying specific features of the register-controlled corpora. They thus serve as a basis of comparison (cf. Neumann 2003) and are annotated with the same features as the register-controlled corpora. At present, each of the reference corpora contains 1,000 word samples from 15 registers and is built roughly following the FLOB corpus design (Hundt et al. 1998). The reference corpora are currently expanded to 2,000 words per register, and two additional registers are included. While we compare realisations of the analysed features in the register-controlled corpora against the background of the reference corpora, we use large corpora like the British National Corpus (http://www.natcorp.ox.ac.uk/) and the Digital Dictionary of the 20th Century German Language (*Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts*, http://www.dwds.de) for large-scale string-based and part of speech comparisons.

The register-controlled corpora (EO and GO) comprise the 8 registers discussed above in both languages. EO and GO are therefore cover terms which include the subcorpora of EO_popsci, EO_fiction etc. The registers are selected because they are relevant for translation – our main object of research. Thus, the translation corpora (ETrans and GTrans) represent the same registers as the EO and GO sub-corpora, but contain translated texts in both directions. The texts in ETrans and GTrans are translations of the texts in EO and GO. The CroCo Corpus thus comprises

- multilingually comparable texts (ER and GR, EO and GO),
- monolingually comparable texts (EO and ETrans, GO and GTrans) and
- parallel texts (EO and GTrans, GO and ETrans).

All in all, the CroCo Corpus as illustrated in Figure 2 covers a basis of comparison for the investigated languages and registers as well as translations and originals. It will be expanded to comprise 1 million words in the course of the project (not including the reference corpora).

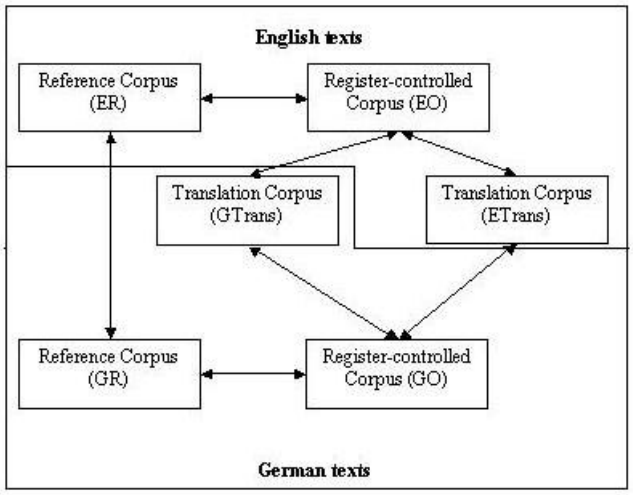

Figure 2. The CroCo Corpus design

We have shown in Section 2 that linguistic annotation of the analysed corpus is required for investigating a linguistically motivated notion of explicitation. In the following section, we will discuss how this annotation is conceived in CroCo.

4 Corpus annotation and mark-up

4.1 Corpus representation

The corpus files in the CroCo project are encoded in XML on the basis of the Text Encoding Initiative (TEI) standard and employing a standard XML editor. This guarantees the exchangeability and searchability of the corpus and its mark-up. The text body is annotated for headings, sentences, paragraphs etc. Each text is encoded in terms of a header that provides meta-information on title, author, publication, register information, etc (see also Section 3). In case of a translation corpus, we encode information on the translator, the translation, the translation process, the author of the original and the source language text. This information is important to filter the corpus and divide it into sub-corpora. Moreover, it enables the corpus user to investigate the corpus on the basis of register information or other independent variables. Thus, investigations concerning specific translation directions, publication dates, nationalities, etc. are possible.

For the analysis of explicitation in CroCo, the different languages and annotation layers require the application of several corpus analysis tools. Each of the tools used for encoding, annotation and exploitation employs different input formats that do not necessarily match straightforwardly. Some tools operate on raw text or tokenised corpora, others on TSV format and others on SGML or XML. Moreover, the outputs that are generated from coding are different across tools. Part of this problem can be dealt with simply by format transformations (using a Perl script or XSLT, i.e. XML transformations on the basis of style sheets; cf. Teich & Hansen 2001). Moreover, a uniform representation of the multi-layer annotation based on an XML stand-off mark-up and a document type definition (DTD) is used to ensure the exchangeability and usability of the corpus (cf. Hansen 2003). Stand-off mark-up means that the different layers of linguistic annotation are kept in separate XML files (see Figure 3), which allows the annotation of overlapping segments where no dominance relation is involved (Teich et al. 2001). This kind of XML encoding takes advantage of the rich set of tools readily available to edit, validate, transform and query the linguistic annotation. The annotated elements represented in independent files are linked through a unique ID. This allows the alignment of source and target language text on the one hand and queries across different annotation layers on the other. In the DTD, which can be seen as a formal grammar for annotation, the units of annotation (i.e. the elements) and their attributes are defined. This

helps to control the error-prone process of manual and semi-automatic annotation, since the corpus annotation can always be validated against its DTD.

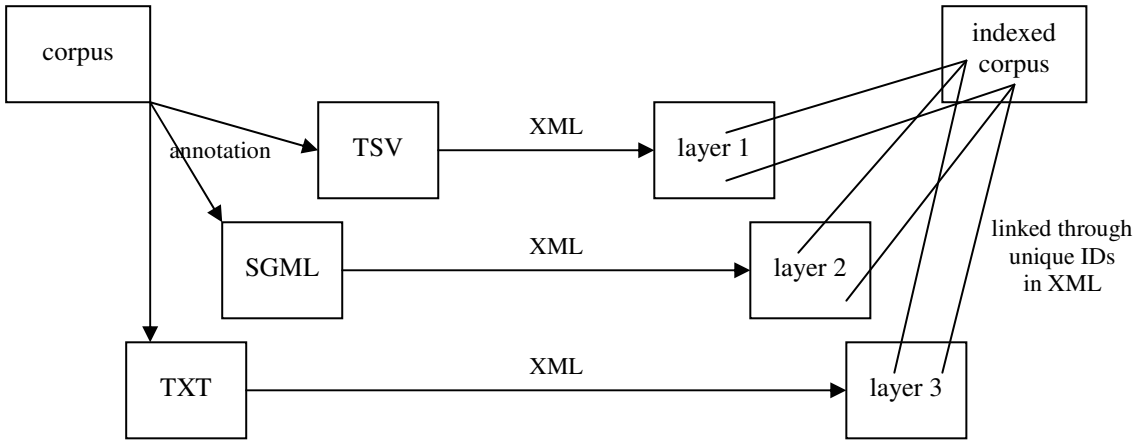

Figure 3. Multi-layer annotation in XML stand-off mark-up

4.2 Alignment

For the analysis of a parallel corpus, the units of translation (i.e. source language text units and their translational equivalents) need to be aligned (see Figure 4). There are various alignment programs freely available and aligners are often incorporated in translation memories. The most commonly implemented technique is sentence-by-sentence alignment. In CroCo, however, we also need to investigate and compare smaller translation units. For this purpose, we develop an algorithm which enables the alignment of clauses, phrases as well as words. The aligned words, phrases, clauses and sentences are stored in separate files, linking the sentences through XML IDs and IDREFs (see Figure 6 for examples).

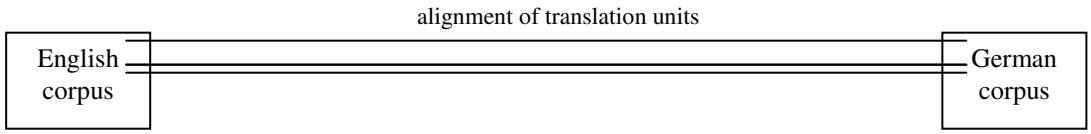

Figure 4. Alignment of source and target text

4.3 Linguistic annotation

Depending on the lexico-grammatical feature on the basis of which explicitation is analysed, annotation is carried out automatically, semi-automatically or manually. In the following, we illustrate an annotation which requires automatic, semi-automatic and manual treatment: Explicitation can easily be identified through nouns and their pre- and postmodifications. A finite relative clause, for example, is more explicit than a non-finite one and a non-finite relative clause can again be more explicit than a prepositional phrase as shown in Example 3. Here, (a) contains information on what the girl is doing as well as on tense. In (b), tense information is implicit, while in (c) even the lexical information of the verb is implicit.

(a) the girl who was standing in the corner
(b) the girl standing in the corner
(c) the girl in the corner

Example 3. Taken from Quirk et al. (1985:1243)

Quirk et al. (1985), however, also state that a premodifying structure might reduce explicitness compared to a postmodifying structure. In version (a) of Example 4 the information on the road direction is implicit.

(a) the Lincoln road
(b) the road to Lincoln
Example 4. Taken from Quirk et al. (1985:1330)

In consequence, nouns and their modifications are annotated. For this purpose, an automatic part-of-speech tagging is applied. On the basis of this part-of-speech tagging, the nouns can automatically be identified as heads with the help of a Perl script. In a further step, adjectives, adverbs, determiners and other one-word modifiers can be detected – again with a Perl script. However, this process is supervised by a human annotator since it is too error-prone to be done fully automatically. In a final step, the human annotator manually assigns the correct post- and premodification tags to the more complex modifiers. For this annotation an XML editor is used, which automatically checks the syntactic correctness of the encoding. The annotation guidelines are reflected in the DTD, which ensures the validity of the encoded translation units (here, the term *translation unit* is used to describe bilingual segmentation units which incorporate a translation relation).

The differences in the languages involved make it necessary to develop two different annotation schemes for the English and the German sub-corpora separately (cf. Hansen-Schirra & Neumann 2003). However, existing commonalities, which can be found in the system of the source as well as in the system of the target language, are preserved in the annotation schemes. Contrasts in the lexico-grammar of the languages involved are compared by basing the annotation scheme on the functions the differing features serve. Thus, the annotated units are comparable across languages.

In each layer, different units are annotated (phrases, words, clauses, etc.). In the translation corpus, they are linked through IDs and IDREFs. A translation unit can comprise several annotation elements. For instance, a premodification tag and a postmodificaiton tag can be assigned to the same noun phrase. Thus, not only can translation units be compared, but also the annotation assigned to the translation units is comparable across languages. This means that a new representation standard is developed which allows the alignment of annotated translation units (see Figure 5).

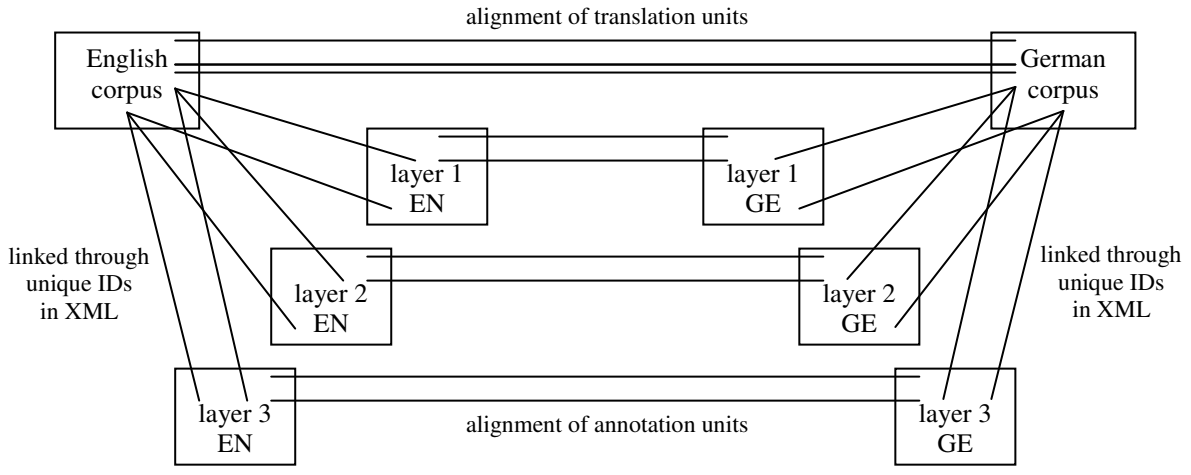

Figure 5. Alignment of source and target language annotation

Figure 5 shows that the different annotation layers in the translation corpus are kept separate using XML stand-off mark-up. The annotation units are linked through IDs and IDREFs. This representation standard developed in the CroCo project for the annotation of translation corpora allows not only the alignment of raw text sequences but also of annotation units. This facilitates the exploitation of the corpora since the extraction of corresponding translation

units as well as their annotation units is a straightforward step in the querying process. On the basis of simple XSLT queries, text segments, translation units and/or annotation units of several annotation layers can be extracted and compared.

An example of the CroCo representation standard is shown in Figure 6. Here, the English noun phrase "nondiscriminatory trade" is translated into German with the noun phrase "Gleichbehandlung im Handel" (*equal-treatment in trade*). The English head "trade" is premodified through the adjective "non-discriminatory". In the German version, the new head of the noun phrase is "Gleichbehandlung" (*equal-treatment*) which is postmodified by the prepositional phrase "im Handel" (*in trade*), where "Handel" (*trade*) is the head of the prepositional phrase. Here, the translation is in fact more explicit than the source language segment.

```
<tu lang="en" id="en-8" idref="ge-8">
        <pre-modification type="adjective">nondiscriminatory</pre-modification>
                <head id="en-22" idref="ge-21">trade</head>
</tu>

<tu lang="ge" id="ge-8" idref="en-8">
        <head id="ge-20" idref="" transfeat="changed">Gleichbehandlung</head>
                <post-modification type="prepositional-phrase">
                        <pre-modification type="determiner">im</pre-modification>
                                <head id="ge-21" idref="en-22" transfeat="changed">Handel</head>
                </post-modification>
</tu>
```

Figure 6. CroCo representation standard[2]

As can be seen in Figure 6, the translation units (*tu*) refer to each other through IDs and IDREFs. This reflects the alignment of source and target language text (in this case, phrases are aligned). In addition, a language attribute is assigned to the translation unit. Postmodifications as well as premodifications of nouns are annotated as elements. The kind of modification is encoded in the attribute *type*. Nouns carry the element *head* in their tag. The corresponding heads in the source and target language segments are also linked through IDs and IDREFs. The alignment of annotation facilitates the comparison of annotation units across languages (here, the comparison of English and German head nouns). Another facilitation is the annotation of the attribute translation feature (*transfeat*) which indicates whether or not the modification of a head is changed. On the basis of this attribute, changed modifiers can automatically be extracted for further investigation.

Figures 5 and 6 show how annotations across layers as well as annotations across languages refer to each other in the CroCo Corpus. The alignment of translation units as well as annotation units on the basis of XML stand-off mark-up guarantees the exploitability of the corpus and makes it a unique resource in translation studies.

5 Extensions in the future

After completing the CroCo Corpus as described in Section 3 we will elaborate the queries needed to investigate explication empirically in terms of features on the levels of lexis, grammar and semantics. Word count methods are employed to get a first idea of explicitation, for instance, by calculating the relation between content and function words, the number of words per text/sentence/clause/phrase, the higher relation of verbal vs. nominal constructions or the higher relation of conjunctions vs. prepositions, verbs vs. nouns, adverbs vs. adjectives, finite verbs vs. non-finite constructions (more explicit vs. more condensed text). In the deeper

[2] To improve the readability of this example, the actual words are displayed instead of the word IDs required for XML stand-off mark-up. In the CroCo Corpus, the English and German version are of course kept in separate files.

analysis, new constituents are identified and more explicit lexico-grammatical realisations serve as indicators. For this deeper analysis, an annotation scheme is developed on the basis of which the corpus is annotated automatically, semi-automatically as well as manually. The annotation covers part of speech tagging, phrase chunking, morphological and sense tagging as well as the alignment of text and annotation as described in Section 4.3. This will result in a bilingual treebank with parallel annotation.

For the empirical analysis of the CroCo Corpus, KWIC indices will be used for word-based searches, TigerSearch for syntactic investigations and XSLT for multi-layer queries. Last but not least, our goal is to make parts of the CroCo Corpus and its annotation available for the public.

References

Baker, M. (1993) Corpus Linguistics and Translation Studies. Implications and Applications, in M. Baker, G. Francis, E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair* (Amsterdam, Philadelphia: Benjamins), 233-250.

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), 175-186.

Biber, D. (1990) Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. *Literary and Linguistic Computing* 5/3, 257-269.

Biber, D. (1993) Representativeness in Corpus Design *Literary and Linguistic Computing* 8/4, 243-257.

Blum-Kulka, S. (1986) Shifts of cohesion and coherence in Translation, in J. House and S. Blum-Kulka (eds.) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies* (Tübingen: Gunter Narr), 17-35.

Frankenberg-Garcia, A. (2004) Are translations longer than source texts? A corpus-based study of explicitation. Paper presented at the Third International CULT (Corpus Use and Learning to Translate) Conference, Barcelona, 22-24 January 2004. Available from http://www.linguateca.pt/Repositorio/Frankenberg-Garcia2004.doc (accessed June 16th, 2005)

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

Hansen, S. (2003) *The Nature of Translated Text – An Interdisciplinary Methodolgy for the Investigation of the Specific Properties of Translations*, Saarbrücken Dissertations in Computational Linguistics and Language Technology 13 (Saarbrücken).

Hansen-Schirra, S. and Neumann, S. (2003) The challenge of working with multilingual corpora, in S. Neumann and S. Hansen-Schirra (eds.) *Proceedings of "Multilingual Corpora: Linguistic Requirements and Technical Perspectives"*. Pre-Conference Workshop at Corpus Linguistics 2003, March 27, 2003, Lancaster, UK, 1-6.

Hundt, M., Sand, A., Siemund, R. (1998) *Manual of Information to accompany the Freiburg - LOB Corpus of British English ('FLOB')* (Freiburg: Albert-Ludwigs-Universität Freiburg).

Neumann, S. (2003) Exploitation of an SFL-annotated multilingual register corpus, in A. Abeillé, S. Hansen-Schirra, H. Uszkoreit (eds.): *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03).* Budapest, 85-92.

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

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985) *A Comprehensive Grammar of the English Language* (London: Longman).

Steiner, E. (2001a) Translations English – German: investigating the relative importance of systemic contrasts and of the text type "translation" *SPRIKreports* 7 (2001), 1-49.

Steiner, E. (2001b) Intralingual and interlingual versions of a text – how specific is the notion of *translation*? In E. Steiner and C. Yallop (eds.) *Exploring Translation and Multilingual Text Production: Beyond Content* (Berlin, New York: Mouton de Gruyter), 161-190.
Teich, E. (2003) *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts* (Berlin and New York: Mouton de Gruyter).
Teich, E., Hansen, S. (2001) Towards an integrated representation of multiple layers of linguistic annotation in multilingual corpora. *Online-Proceedings of Computing Arts 2001: Digital Resources for Research in the Humanities* (Sydney).
Teich, E., Hansen, S., Fankhauser, P. (2001) Representing and Querying Multi-layer Annotated Corpora, in *Proceedings of the IRCS Workshop on Linguistic Databases* (Philadelphia), 228-237.
Toury, G. (1995) *Descriptive Translation Studies and Beyond* (Amsterdam, Philadelphia: Benjamins).
Vinay, J.-P. and Darbelnet, J. (1995) *Comparative Stylistics of French and English. A methodology for translation.* (Amsterdam, Philadelphia: Benjamins). Translation of Vinay, J.-P. and Darbelnet, J. (1958) *Stylistique compare du français et de l'anglais.*