CROCO

LINGUISTIC PROPERTIES OF TRANSLATIONS
A CORPUS-BASED INVESTIGATION FOR THE LANGUAGE PAIR ENGLISH-GERMAN

Title **Categories for the Annotation**

Author Stella Neumann, Silvia Hansen-Schirra

Deliverable No. 4
Work package 1.5, 1.6
Status *final version*
Availability *public*
Date *26 June 2006*

DFG project STE 840/5-1
http://fr46.uni-saarland.de/croco/

Categories for the Annotation

1 Introduction

The aim of the CroCo project is to learn more about the assumed property of explicitation in translations. The object of investigation is the corpus as described in Deliverable no. 1. We expect to gain insight into explicitation from analysing the linguistically enriched corpus. The underlying principle of the investigation is to distinguish between the linguistic annotation and the interpretative derivation of indicators and query of the enriched corpus in view of our research question. This means that the categories for the annotation contain as little interpretation as possible. This principle will be described in more detail in section 2 below.

We, thus, only annotate (lexico-grammatical) categories which are expected to serve as indicators for explicitation (cf. Steiner 2005). These categories comprise parts of speech and morphology for terminal nodes as well as phrase structure and grammatical functions for the highest nodes in the sentence. All of these are annotated and/or processed electronically. For parts of speech and morphology existing tools are employed using their annotation schemes. The main aim of the present paper is to describe the categories chosen as well as the tools used to annotate the CroCo corpus.

2 Distinguishing between annotation and interpretation

As previously mentioned, the guiding principle in the process of enriching the CroCo corpus with linguistic information is the clear distinction between lexico-grammatical categories and interpretation of indicators. This distinction allows a clean methodological procedure starting from a hypothesis of the type “If a text is a translation, it will display explicitation”. We can replace “explicitation” in this hypothesis with any other phenomenon to be investigated in originals and translations. In a next step we can operationalise this hypothesis with variables and find observable indicators which can reasonably be assumed to measure the variables. This procedure offers a twofold benefit. First, as mentioned, we work in a methodologically clean way, secondly, and more importantly, the corpus becomes a resource that can be used for the analysis of other research questions as well since the annotation is not specifically geared towards explicitation. Once the resource is enriched with comprehensive linguistic information and aligned on different annotation layers, a variety of queries is possible. The queries that will retrieve findings on explicitation will be described in a later deliverable. In the following we will give examples of potential information on explicitation to be expected from the annotated features wherever possible.

3 Technical aspects of multi-layer annotation

When working with several annotation tools, particularly on multiple layers, the differing output formats are a classical problem. We deal with this by using an XML stand-off format for all annotations. This means that we have to convert all annotation outputs to a uniform XML version. As the units on the different layers may overlap, each annotation is stored in a separate file. For a more detailed description of the technical realisation of the CroCo annotation see Hansen-Schirra et al. 2006.

4 Multilevel annotation

The annotation of the CroCo Corpus covers the following layers: parts of speech and morphology on the word level as well as phrasal categories and grammatical functions on the phrase level. The categories of annotation and the annotation tools are described in the following. Operationalisation rules and examples can be found in the annotation guidelines which are part of the corpus documentation and which will be offered with the corpus release.

4.1 Parts of speech

Part of speech tagging means that each token in a corpus receives a tag categorizing its word class. A corpus enriched with part of speech information offers a range of possible interpretations. For instance, we can retrieve information on the amount of nominal versus verbal elements in the corpus[1]. Another example is filtering the annotated corpus for open class and closed class words. This filter can then be used to calculate lexical density in the corpus. Both examples offer information that can be interpreted in view of explicitation.

For part-of-speech tagging and tokenisation, we use the TnT-Tagger (Brants 2000), a dedicated tool assigning detailed tags on the basis of a statistical calculation. TnT can easily be trained on different languages. We use existing tag sets for English and German, the languages analysed in CroCo, i.e. the Susanne tag set for English (Sampson 1995) and the Stuttgart-Tübingen Tag Set for German (STTS; Schiller et al. 1999). Minor mistakes are corrected by writing an entry into the tool's lexicon. TnT's output format is TSV which can be transformed easily to XML.

4.2 Morphology

Information on morphology is particularly of interest in German but is also annotated in the English subcorpora. It is interesting in view of explicitation by itself, for instance when an English non-finite construction is translated by a German finite construction requiring morphological marking of tense, mood, voice etc. Lemmatising the corpus allows the calculation of the type-token ratio. Additionally, morphological annotation also serves as a building block for other annotations as well as complex queries which can then be interpreted with respect to explicitation.

The tool used for this annotation step is MPRO (Maas 1996). It does not only provide information on morphology and lemmatisation but also on phrase chunking (see section 4.3), some hints on semantics as well as tokenisation. Furthermore, MPRO also contains part of speech tagging. However, it is less fine-grained than the tagging provided by TnT (see section 4.1) and is therefore not used.

MPRO is rule-based, its annotation scheme thus containing a list of grammar rules. It works both in English and German (as well as a number of other languages; see Maas 1998). MPRO's output format is a structured text format. In order to make it searchable in combination with the other annotations in CroCo it is transformed into XML.

4.3 Phrase chunking and grammatical functions

The level of phrases is approached in two ways. First, we use the MPRO output to obtain an automatic phrase chunking. Additionally, we analyse grammatical functions manually. This manual analysis also includes a formal phrase structure analysis complementing the error-prone automatic output. In the manual analysis, we concentrate on the highest nodes in the sentence structure. Each phrase chunk is annotated for type of phrase and grammatical function.

Phrases may offer information on explicitation in certain constellations. For instance, when the words of a phrasal object in one text can be linked to the words of a clausal object in the correspondent text, this can be an indicator for explicitation or implicitation (depending on the direction of the link). Querying phrases may therefore be useful in combination with the alignment (see section 5). The mismatch of word alignment against the annotation of grammatical functions, a so-called “crossing line”, may be interpreted in view of other

[1] This information is best contrasted to a basis of comparison, for instance by comparing the amount in a register-controlled corpus to the amount in a register-neutral corpus (cf. Deliverable no. 1 and Neumann 2003).

properties of translation as well.

The grammatical functions are annotated in a theory-neutral fashion. The German annotation scheme for both phrase structure and grammatical functions is based on the TIGER[2] annotation scheme with some changes where appropriate. These build on descriptions in Helbig & Buscha (2001) and the Duden Grammar for German (Duden 1998). The English annotation scheme is based on Quirk et al. (1985). The categories are the following: subject, direct object, indirect object, genitive object (for German only), prepositional object, complement, finite, predicator, adverbial, conjunction, apposition, minor clause, particle and negation. The guidelines for both languages are constantly reviewed during the manual annotation process.

For the automatic phrase chunking the MPRO output as described in section 4.2 is converted in the CroCo format. The manual annotation is done using MMAX2 (Müller & Strube 2003), a tool that allows marking multiply nested annotation units as well as linking several units. This latter feature is used in CroCo for aligning clauses. The pre-processing necessary for the preparation of the corpus before using MMAX is described in Deliverable no. 3. This deliverable also describes how to align the clauses in the CroCo Corpus with MMAX2. Since the annotation of grammatical functions is – from a technical perspective – done in a similar way, Deliverable no. 3 can be used as a short manual for the CroCo annotation and alignment with MMAX2. MMAX produces an XML output which has to be adapted to the CroCo format.

5 Multilevel alignment

This special kind of annotation is of utmost importance for the interpretation of translation properties. It is only on the basis of the alignment that the multilevel annotation becomes meaningful. The distinctive characteristic of the CroCo alignment is that it is not restricted to one level, e.g. word level which is typically used in machine translation or sentence level which is typically used with translation memories, but is implemented on three different levels: word level, clause level and finally sentence level. This layout has far-reaching consequences for the analysis of the annotated corpus: if the alignment within a given segment on the different layers is not parallel but crosses lines, this represents a finding that will probably be relevant for the interpretation of translation properties. Another finding that will be of interest concerns empty links on any of the alignment layers (cf. Hansen-Schirra et al. 2006). Figure 1 below shows an example of a sentence containing both crossing lines and empty links. The alignment on the different layers is carried out with existing tools where tools are available.

[2] http://www.ims.uni-stuttgart.de/projekte/TIGER/

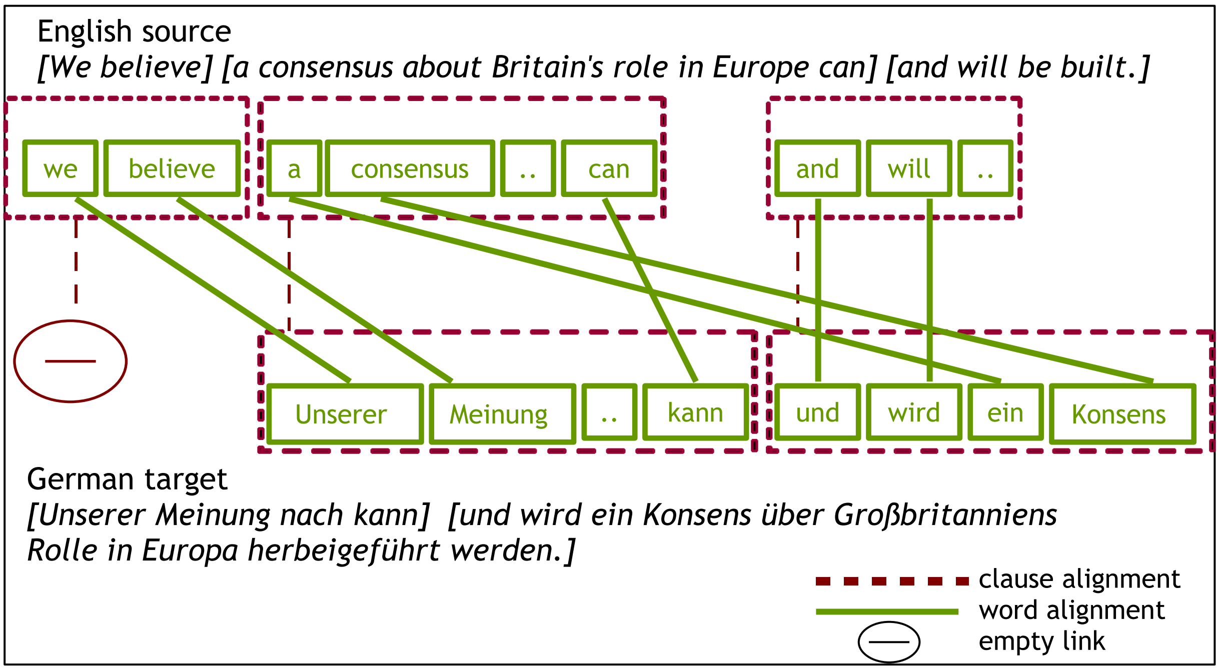
English source
[We believe] [a consensus about Britain's role in Europe can] [and will be built.]
we
believe
a
consensus
..
can
and
will
..
Unserer
Meinung
..
kann
und
wird
ein
Konsens
German target
[Unserer Meinung nach kann] [und wird ein Konsens über Großbritanniens Rolle in Europa herbeigeführt werden.]
clause alignment
word alignment
empty link

Figure 1. Example of crossing lines and empty links in two alignment layers

The different alignment layers and the alignment tools are described in the following. Operationalisation rules and examples can be found in the annotation guidelines which are part of the corpus documentation and which will be offered with the corpus release. A more technical description of the multilevel alignment is given in Deliverable no. 3. In this deliverable, the technical realisation of the alignment process including the necessary pre-processing steps is described.

5.1 Word alignment

As word aligned corpora are used as a resource in statistical machine translation, methods and tools for word alignment typically come from this field of research. In CroCo, we use a new alignment technique which makes use of explicitly structured information (cf. Schrader 2006). This means that multiple annotation layers (e.g. parts of speech, morphology, lemma, syntactic information) are used to improve the quality of the alignment output. Moreover, this alignment tool combines statistic calculations with linguistic input. We chose it because it produces acceptable results if trained on a large corpus. In order to improve the results, we will train the tool not only on the CroCo corpus (taking into account the different registers and annotation layers) which is relatively small for this purpose but also on the German-English part of the Europarl Corpus (Koehn 2005). Due to the limited budget of the project and the costly remaining annotation the word alignment will not be revised manually. For further processing in CroCo the output is transformed into XML.

5.2 Chunk alignment

It is not necessary to run an alignment procedure on the chunk level. Phrase alignment can be derived from word alignment in combination with the phrase chunking described in section 4.3. Syntactic functions can be mapped automatically across the parallel corpus, since the functional units apply to both languages.

In the translation process, however, information is often moved from one grammatical function to the other, thus constituting a typical feature of translation, potentially with an impact on explicitness. We can retrieve these units on the basis of the word alignment and the annotation of grammatical functions. A possible query for this phenomenon would be: Return all units which are aligned on the word level and which belong to a subject in the source text but to an object in the target text. With this query, all subject-to-object shifts in translations can be retrieved.

5.3 Clause alignment

Within the CroCo project, clauses are aligned according to their semantic contents. Finite and non-finite verbs serve as a basis for segmenting and aligning clauses. This means that a clause consists of one finite or non-finite verb (excluding central modals, which are not separated from their full verbs) as well as the corresponding constituents. For a detailed description of the boundaries between verbal and nominal as well as verbal and adjectival constructions see the annotation guidelines which are part of the corpus documentation and which will be offered with the corpus release.

For the alignment of clauses in CroCo, again MMAX2 is used (Müller & Strube 2003). For a more detailed description of the pre-processing steps as well as the alignment process in MMAX see Deliverable no. 3. MMAX produces an XML output which has to be adapted to the CroCo format.

5.4 Sentence alignment

Translation memories typically build on sentence alignment. The common tools all contain a sentence alignment tool. The one used for the CroCo alignment is WinAlign from the Trados Translator's Workbench[3]. This tool produces fairly good results with the remaining mistakes being revised by the annotator in a user-friendly GUI.

Segmenting is based on a pragmatic definition of the sentence like the following: "A sentence is a syntactically autonomous sequence of words, terminated by a full-stop punctuation" (Simard 1998). The decision whether to align two segments is made by human interpretation. The possible alignments are limited by the restrictions of WinAlign. It does not allow linking more than one to two or more segments. In the case of empty links, the annotator has to insert an empty segment, because otherwise the non-aligned segment will be lost in the output. The output is a plain text file with each original sentence and its translation in one line divided by a semicolon. For further processing in CroCo it is transformed into XML.

6 Outlook on the queries

So far, we have only described some superficial interpretations of the annotations. The interpretation has to be based on queries into the annotations. The result of the queries still does not answer our research question, but has to be subjected to thorough interpretation before being able to make any statements in relation to explicitation or any other property of translation. It may have become clear from the above description that the queries have to extend to several layers of annotation in order to produce useful results. A first outlook on how this may look like is given in Hansen-Schirra et al. (2006).

References

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.

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

Deliverable no. 1. The CroCo Corpus design. CroCo Linguistic Properties of Translations. Saarland University, Saarbrücken.

Deliverable no. 3. Multidimensionales Korpus-Alignment. CroCo Linguistic Properties of Translations. Saarland

[3] http://www.trados.com/

University, Saarbrücken.

Duden Grammatik (1998) *Der Duden: Grammatik. der deutschen Gegenwartssprache*. Mannheim.

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

Hansen-Schirra, S., Neumann, S. & Vela, M. (2006) Multi-dimensional Annotation and Alignment in an English-German Translation Corpus. In: *Proceedings of the 5th Workshop on Multi-dimensional markup in NLP*. Trento. 35-42.

Helbig, G. & Buscha, J. (2001): *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht.* Berlin/München: Langenscheidt.

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).

Koehn, P. (2005) Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of MT Summit X.*

Kuhn, J. (2004) Experiments in parallel-text based grammar induction. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics: ACL 2004*, 470-477.

Maas, H.-D. (1996) MPRO – Ein System zur Analyse und Synthese deutscher Wörter. In Hausser R. (Hrsg.) *Linguistische Verifikation, Sprache und Information. Dokumentation zur Ersten Morpholympics 1994*. Tübingen: Niemeyer, 141–166.

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

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

Neumann, S. (2003) *Textsorten und Übersetzen. Eine Korpusanalyse englischer und deutscher Reiseführer.* Frankfurt/M. u.a.: Peter Lang.

Och, F. J. & Ney, H. (2000) Improved Statistical Alignment Models *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hongkong, China, October 2000, 440-447.

Quirk, R.; Greenbaum, S.; Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Sampson, G. (1995) *English for the Computer*. Oxford: Clarendon Press.

Schiller, A., Teufel, S., Stöckert, C., Thielen, C. (1999) *Guidelines für das Tagging deutscher Textcorpora mit STTS.* Stuttgart: Universität Stuttgart.

Schrader, B. (2006) How does morphological complexity translate? A cross-linguistic case study for word alignment. *Proceedings of Linguistic Evidence Conference*, Tübingen: 189-191.

Simard, M. (1998) The BAF: A Corpus of English-French Bitext. *Proceedings of LREC 1998.* http://www.iro.umontreal.ca/~simardm/lrec98/

Steiner, E. (2001) 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.

Steiner, E. (2005) Explicitation, its lexiocogrammatical realization, and its determining (independent) variables – towards an empirical and corpus-based methodology. *SPRIK-reports* no. 36. Oslo.