CROCO

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

Title	***Corpus Design***
Author	Stella Neumann

Deliverable No. 1
Status *Final*
Availability *Public*
Date *October 26, 2005*

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

The CroCo Corpus design

1 Introduction

This paper describes major aspects of the corpus design for the CroCo project. The CroCo Corpus is conceived as a resource capable of investigating a wealth of research questions concerning the specificities of translated texts as compared with non-translated, i.e. original, texts in the language pair English-German. This has consequences for its structure, its size, the selection of texts etc.

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. Although the CroCo project concentrates on explicitation the corpus will permit looking into the other translation properties like simplification, normalisation, levelling out and shining through. Finally, the need to explain why translations differ from originals adds another dimension to the criteria for the corpus design. In what follows we describe the resulting decisions for the CroCo Corpus.

2 Structure of the corpus

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 2,000 words from 17 registers and is built roughly following the FLOB corpus design (Hundt et al. 1998). Each 2,000 word sample is again subdivided into approximately 6 samples taken from different texts by different authors. The registers are *press reportage, editorial, review, religion, skills, popular lore, biographies, political texts, science, general fiction, mystery, prepared speech, cooking recipe, romance, call for tender, travel guide book* and *court decision.*

The register-controlled original corpora (EO and GO) comprise 8 registers discussed below in both languages. EO and GO are therefore cover terms which include the subcorpora for each of the 8 registers. We have selected these registers 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 GO and EO. 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 1 covers translations and originals as well as a basis of comparison for the investigated languages and registers. It will be expanded to comprise 1 million words in the course of the project (not including the reference corpora).

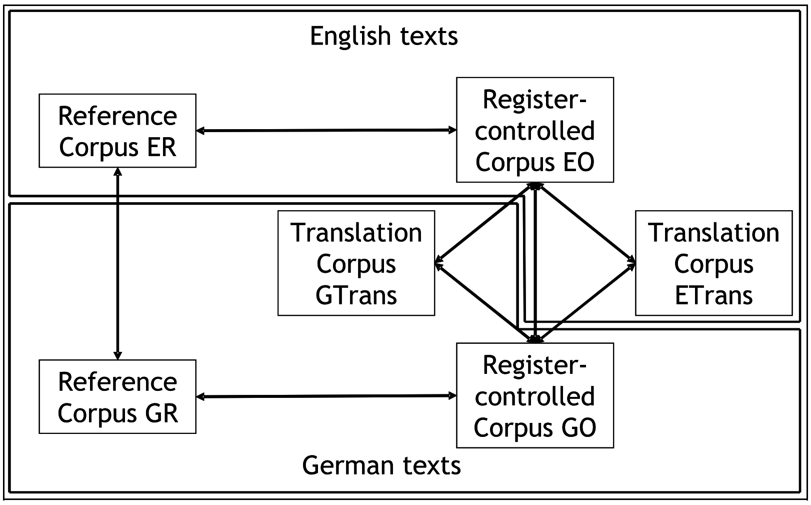
English texts
Reference Corpus ER
Register-controlled Corpus EO
Translation Corpus GTrans
Translation Corpus ETrans
Reference Corpus GR
Register-controlled Corpus GO
German texts

Figure 1. The CroCo Corpus design

3 Design criteria

The design criteria we addressed are representativeness, balance and comparability and in connection with this latter criterion also register. Representativeness will be discussed in section 4 below. 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, 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 the year 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 2001, 2004 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 is of special interest in at least one register included in the corpus. This resulted in the decision to include the 8 registers listed in Table 1.

Two additional registers are included in 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 (see section 5). This allows us to filter the corpus according to specific register features.

For the reference corpora we relied on the design of the FLOB corpus but replaced three registers, because FLOB seems somewhat biased towards fictional texts. We thus removed *science fiction, adventure and western* and *humour.* Additionally, we added the two registers *travel guide book* and *court decision* to the reference corpora.

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

4 Corpus size

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. For instance, we can count all people living on a given stretch of earth, but 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. 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. A smaller corpus which is richly annotated is preferable to a large one without much annotation which may not add any information relevant to the research question. We therefore follow Douglas Biber (1990, 1993) who shows that smaller corpora – if well-balanced – are capable of covering all linguistic features of a given register. His calculations, i.e. 10 texts per register with a length of 1,000 words, serve as an orientation for the size of our core corpus.

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. 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 analysis. Needless to say that text length may not be the only criterion for a sub-corpus taken from the CroCo Corpus.

Given the 8 registers as well as the targeted corpus size of 1 million words in the four sub-corpora EO, GO, GTrans and ETrans (excluding the reference corpora) we obtain the following calculative size of the register samples.

Each sub-corpus contains 250,000 words. Divided into the 8 registers, this adds up to a calculative size of 31,250 words per register sample. Each sample comprises at least 10 texts resulting in a computed length of 3,125 words per text. In some of the registers, texts are typically shorter. In these cases, we increase the number of texts in order to obtain the targeted size of the register sample.

5 Text sampling

The 10 or more texts per register are selected in a quota sample[1]. Ideally, the texts would be selected on the basis of a random sample (the worst selection method being drawing the next best texts). Random sampling requires a determinable population, i.e. the number of all texts that are associated with the given register (assuming that it is possible to delimit the register in such a way as to allow distinctly associating each appropriate text with it). However, the population cannot be determined with almost all registers. Therefore, we have to approximate criteria differentiating elements representing the given register. These may be author, publisher, subject matter, language variety, etc. depending on the requirements of the respective register. The preferred sampling method is thus quota sampling, i.e. selection not randomly but according to a fixed quota, here e.g. one text per author etc.

This does not mean that we exclude random sampling as such. It comes into play in two cases. First, we partly make use of existing corpora like the FLOB corpus for our English reference corpus (ER), as mentioned previously. We take samples from FLOB for some registers with the help of automatic calculation of random numbers. Secondly, in cases where we draw samples from longer texts (see section 4) we choose the samples by calculating random page numbers.

The complete information on each text together with a brief register analysis is kept in a file containing the metadata for each text. The original file can be traced back by its file name contained in the metadata. The complete set of metadata, which is based on the TEI standard, is displayed in Figure 2.

```
<teiHeader>
      <fileDesc>
            <filename/>
            <subcorpus/>
            <language/>
            <titleStmt>
                  <title/>
                  <author/>
            </titleStmt>
            <translation/>
            <publicationStmt>
                  <publisher/>
                  <address/>
                  <date/>
                  <distributor/>
                  <availability/>
            </publicationStmt>
            <sourceDesc>
                  <author/>
                  <title/>
            </sourceDesc>
            <registerAnalysis>
                  <register/>
                  <field>
                        <experientialDomain/>
                        <goalOrientation/>
                  </field>
                  <tenor>
```

[1] A quota sample is a nonprobability sample that takes into account the proportion of individual specimens in different population categories within the population.

```
                                <agentiveRole/>
                                <socialRole/>
                                <socialDistance/>
                        </tenor>
                        <mode>
                                <languageRole/>
                                <channel/>
                                <medium/>
                        </mode>
                </registerAnalysis>
        </fileDesc>
        <encodingDesc>
                <projectDesc/>
                <samplingDesc>
                        <extent/>
                        <size/>
                </samplingDesc>
                <profileDesc>
                        <creation/>
                        <langUsage/>
                </profileDesc>
        </encodingDesc>
        <annotation>
                <pos>
                        <respStmt>
                                <name/>
                                <responser/>
                        </respStmt>
                </pos>
                <morph>
                        <respStmt>
                                <name/>
                                <responser/>
                        </respStmt>
                </morph>
                <chunk>
                        <respStmt>
                                <name/>
                                <responser/>
                        </respStmt>
                </chunk>
        </annotation>
</teiHeader>
```

Figure 2. CroCo header

Apart from the above mentioned information, the header additionally contains information on the annotation process. Each step is logged in the "annotation" tag.

6 Relationship with existing corpora

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 (BNC; http://www.natcorp.ox.ac.uk/) and the Digital Dictionary of the 20th Century German Language (DWDS, *Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts*, http://www.dwds.de) for large-scale string-based and part of speech comparisons.
For all more detailed comparisons, we mainly rely on our own reference corpora. This is due to two reasons. First, the registers in our reference corpus are clearer delimited as compared with the large corpora. The BNC, for instance, consists of 10 written text domains which are as broad as "world affairs". The same applies for DWDS with only four categories for written texts. This is problematic because the link between corpus features and indicators of explicitation rests on hypotheses only. The more superficial the features the weaker the link

between indicator and hypothesis will be. Although this is mainly an issue for the formulation of hypotheses, it also means that we have to make sure the corpus data is as clean as possible to avoid interferences stemming from the data. The second reason is that access to both corpora is limited. It is not possible to download the corpus and annotate it with the information we think necessary for the investigation of translation properties. Nevertheless, particularly in the case of concordances, our own reference corpora may constitute too small a sample. Therefore, we may want to run concordances and/or part-of-speech queries on the large national corpora.

The same reasons explained for the reference corpora also apply to the fact that we do not use the German-English part of the Oslo Multilingual Corpus (http://www.hf.uio.no/iba/OMC/) as part of our register-controlled and translation corpora. Furthermore, this resource can only be accessed locally at the participating universities of Oslo and Bergen.

7 Corpus management system

The corpus is stored in the following folder structure created by Mihaela Vela. The Corpus folder is divided into four main folders:

- The **Archive** containing those registers which are only translated in one direction as well as the full texts of the samples drawn for the core corpus where applicable.
- The translation direction **English2German** of the core corpus containing the English originals and matching German translations in the 8 registers described in section 3.
- The translation direction **German2English** of the core corpus containing the German originals and matching English translations in the same 8 registers.
- The **Reference Corpus** containing texts from 17 registers both in English and in German. This corpus is an extension of the reference corpus described in Neumann (2003).

The main folders are then subdivided in folders for each sub-corpus: English Originals (EO), English Translations (ETrans), English Reference (ER), German Originals (GO), German Translations (GTrans), German Reference (GR). On the next level we have folders for each register contained in the respective sub-corpus.

For each register we then discriminate between “Plain”, containing the original texts again subdivided in a folder with the files in its original format (“Source”) and one with the files in .txt-format (“TXT”), “Meta”, containing the matching metadata for each file, and finally “Annotated”, containing the different annotations. This includes the folders “Pos” for part-of-speech-tagged files, “Chunk” for phrase chunked files and “Morph” for the files annotated with morphology. The clear separation of the annotation files is part of the preparation of stand-off XML mark-up.

The file names reflect the format and position of the respective file in the folder structure. Thus, the German original text number 001 in the register WEB for websites would have the file name `GO_WEB_001.txt`, the header file `GO_WEB_001.header`, the matching translation `ETrans_WEB_001.txt`, the file containing the part-of-speech-tagging `GO_WEB_001.tag` and so on. Knowing that this latter file is the PoS-tagged version of the first German original website text, we would look for this file in the path “German2English/GO/WEB/Annotated/POS”.

8 Copyright issues

As of present, copyright for most texts in the corpus is not cleared. One task during the project life time is clearance of at least (meaningful) parts of the resource for a wider public. This will be a cumbersome process as an essential characteristic of the CroCo Corpus is its diversity with respect to authors and publishers.

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.

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

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

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. (2004) The heterogeneity of individual languages as a translation problem. In *Übersetzung – Translation – Traduction.* Ed. by Kittel, H., Frank, A.P., Greiner, N., Hermans, T., Koller, W., Lambert, J., Paul, F.. Volume 1. (Berlin, New York: de Gruyter), 446-454.