How to Overcome Registerial Translation Problems: a corpus-based approach

Silvia Hansen
Mary Klaumann
Stella Neumann
Saarland University, Germany

Este artigo apresenta a aplicação de um corpus no par lingüístico inglês-alemão anotado com base em características especificas de registro. Discute-se a relevância de diferenças translingüísticas de registros para a tradução e mostra-se como tradutores podem se beneficiar de corpora paralelos anotados que sirvam como referências em tempo real oferecendo soluções de tradução orientadas por marcas de registro.

This article aims at outlining an application of an English-German translation corpus annotated on the basis of language specific register features. We discuss the relevance of cross-linguistic register differences for translation and show how translators can benefit from registerially annotated parallel corpora, which serve as on-line references offering register-oriented translation solutions.

Introduction

This article aims at revealing how translation corpora annotated with register-specific features can be used to solve translation problems. Within this context, the importance of this topic for the translator's daily life is discussed in Section 2. Section 3 explains the relevance of cross-linguistic differences exemplified in a given register. In Section 4, the computational techniques which are needed to solve register-specific translation problems in an empirical approach are introduced. Finally, a sample solution for the translation of cross-linguistically diverging register features is presented in Section 5. The article concludes with a summary and some suggestions for future research (Section 6).

Register and translation

According to Steiner (2001:5) the translation process "has to be seen in the context of and in interaction with" typological factors and register features. This paper focuses on the influence of the register in question on the text to be translated. Biber (1995:132) calls registers "text categories readily distinguished by mature speakers of a language". They are realised by "a relatively high or low frequency of occurrence of particular lexico-grammatical features", as Teich (2001:21) puts it.

For translation purposes, first it is necessary to investigate the occurrence or absence of certain features. The linguistic features of one register thus established can then be related to more abstract functions allowing the description of a given register in more general terms. For example, Biber (1995) shows that passive constructions typically occur in English scientific prose on the basis of a multi-dimensional corpus analysis. He establishes a connection between this linguistic feature and the abstract style preferred in scientific writing.

As will be seen later, passive constructions are also a typical feature of German scientific writing. For translation purposes cross-linguistic similarities and divergences of register features have to be identified. More specifically, if – like in the case of passive in scientific prose – a feature has been identified to occur in both languages, the realisations of this feature may still diverge. This holds especially for more abstract grammatical features. The description of contrasts and commonalities and the availability of their realisations in a parallel corpus are crucial for the translator.

A corpus of originals and translations in both languages involved may serve as a database for the translator in order to look for these typical realisations. The information thus provided is useful for the translation of structures which lack an equivalent in the target language and helps finding expressions which are in line with target language norms.

Passive structures

We have taken the genre of scientific prose to illustrate translational problems which may result from registerial variation. The use of passive voice constitutes a major feature of this register in English as well as German (Biber 1995:143; Fluck 1997:92).

Generally speaking it can be said that passive structures focus on the object, effect or the result of an action. Since these foci are of particular interest in technology and science, passive constructions play a major role in the objective description of observations or experiments in English and German scientific writing. This, too, applies to popular-scientific writing, i.e. scientific prose, as one of the sub-categories of scientific writing. The impersonal style achieved by using passive constructions can, however, not be implemented equally in the two languages (Schwanzer 1981:217). Therefore, passive structures seem particularly interesting in connection with the investigation of registerial problems regarding the translation of scientific prose.

By using passive constructions the emphasis shifts from the agent of an action to the recipient, which is often an abstract, inanimate entity. The recipient then becomes the subject of the action whereas the agent, being of minor importance to the description, is often omitted.

The use of passive offers the possibility to avoid the ponderous repetition of the subject if a series of actions is performed by the same person. A connection can also be established between the tendency to organise utterances in theme-rheme sequences (Sager et al. 1980:209, Fluck 1997:119ff), since for example in English, passive voice makes it possible to place the element in question in thematic position, thus emphasising it.

Passive constructions are particularly frequent in English scientific writing due to the lack of alternative impersonal constructions, apart from the rare use of "one" (Sager et al. 1980:209).

German passive constructions in the narrow sense, namely dynamic as well as statal passive resemble the English passive forms. However, in addition to these passive forms in the narrow sense, German uses the following impersonal structures (Fluck 1997:96):

- "man" (one):
  Man gewinnt sie aus Skelettresten
- Support verb constructions:
  Herschbach führte die Methode durch experimentelle und theoretische Beiträge zur Reife
- Infinitive structures (*sein* + *zu* + infinitive):
  Diese gestörte Strahlung ist als Nachricht aus dem Reaktionsknäuel aufzufassen

- *lassen* + *sich* + infinitive
  In der Medizin etwa lassen sich so Trägersysteme konstruieren
- Reflexive verbs
  Tatsächlich fanden sich in der Haut von Fingern und Zehen vielfach noch Zellkerne

These passive alternatives make it possible to present information in an objective yet varied style. This divergence of means offered by the two languages to convey information in an impersonal style has to be dealt with by the translator.

Computational Techniques

In order to overcome these register-specific translation problems, the corpus design displayed in Figure 1 has been used.

The corpus under investigation consists of nine English originals, their translations into German and twelve German originals. The English texts are taken from “Scientific American”, the German ones from “Spektrum der Wissenschaft”. Thus, the register under investigation is scientific prose. On the basis of this corpus design, the parallel sub-corpus (English originals and German translations), the comparable sub-corpus (German originals and German translations) as well as the multilingual sub-corpus (English originals and German originals) can be investigated. Taken together, the corpus comprises 66,177 words.

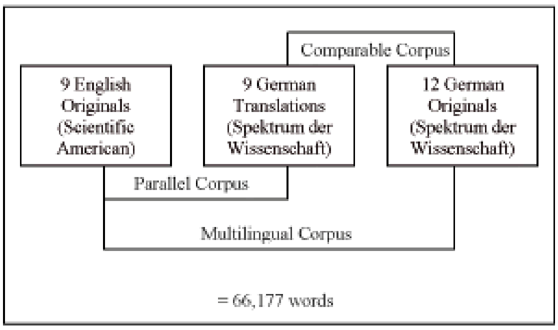

Comparable Corpus
9 English Originals (Scientific American)
9 German Translations (Spektrum der Wissenschaft)
12 German Originals (Spektrum der Wissenschaft)
Parallel Corpus
Multilingual Corpus
= 66,177 words

Figure 1: Corpus design

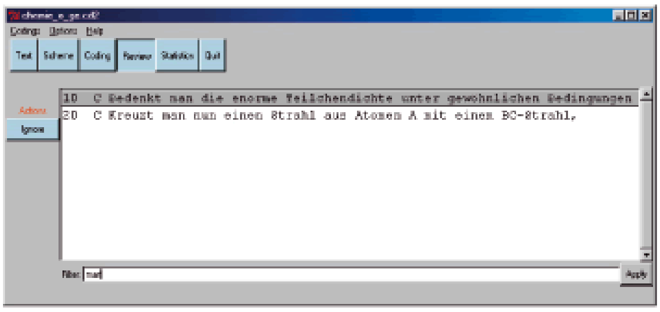
For the annotation of the features, which are typical for the register of popular-scientific texts, i.e., passive in English and passive alternatives in German, the semi-automatic annotation tool Coder (cf. O'Donnell 1995) has been used. This tool allows the segmentation of the corpus into sentences or smaller chunks, the definition of an annotation scheme and the annotation of the segments on the basis of the pre-defined scheme. Changes in the annotation scheme and thus in the annotation itself are supported as well. The output format of the corpora annotated with Coder is SGML (see Figure 2).

Additionally, the system computes a descriptive statistics of the annotated segments. This functionality shows the distribution of active, passive and passive alternatives in the annotated sub-corpora and their importance for English and German respectively. On the basis of this descriptive statistics, the observations concerning the register specificities introduced in Section 3 can be tested for each language (including the comparison of German translations and German originals).

```
<codings>
 <body>
   <segment features="unit" comment="" ignore=1>
         1  The hidden strength of hydrogen
   </segment>
   <segment features="unit active" comment="" ignore=0>
         2  Hydrogen links up with other atoms in many ways,
   </segment>
   <segment features="unit ing-nonfinite-construction ing-
   nonfiniteclause"
    comment="" ignore=0>
         forming a wide variety of compounds, from methane to
      DNA.
   </segment>
 </body>
</codings>
```

Figure 2: Coder's SGML output format

In order to query the corpus, Coder offers a concordance function. Figure 3 shows a concordance for the German passive alternative “man” (one).



```
10  C Bedenkt man die enorme Teilchendichte unter gewohnlichen Bedingungen
20  C Kreuzt man nun einen Strahl aus Atomen A mit einem BC-Strahl,
```

Figure 3: Corpus querying with Coder

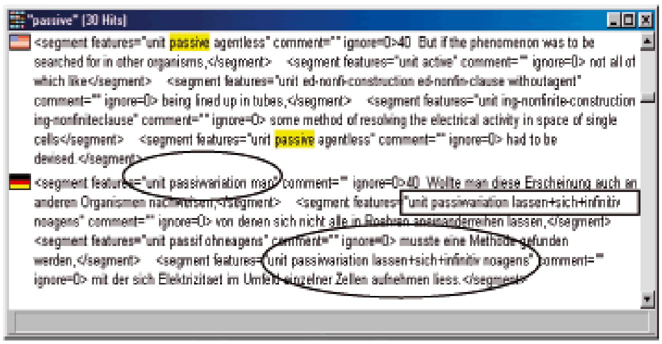
However, corpus annotation and querying with Coder still poses some problems for the investigation of a parallel corpus. For this purpose, the sub-corpus of original texts and the sub-corpus of translations have to be annotated and queried separately. The extraction of annotated segments together with their source or target language equivalents is not possible at all.

For this purpose, we aligned the parallel corpus sentence by sentence with Déjà Vu[1]. Since standard concordance programs for parallel corpora, such as the IMS Corpus Workbench (cf. Christ 1994), allow queries for words and/or part-of-speech tags, but do not support the annotation of more abstract linguistic features which span phrases or even clauses, the parallel corpus has been loaded into a translation memory. This tool is usually applied to the translation of repetitive texts providing a database with bilingual translation equivalents.

The translation memory which has been used for our purposes is the Translator’s Workbench[2]. Since the Translator’s Workbench operates among other input formats on the basis of SGML, Coder’s SGML output format of the annotated parallel corpus (see Figure 2) has been loaded into the Translator’s Workbench. Using this format in combination with the Translator’s Workbench means that the SGML tags are not hidden in the displayed matches, which is not very user-friendly. Nevertheless, this combination of tools and formats allows the querying of words and/or abstract annotations carried out with Coder (see Section 5 for a sample query).

Towards register-oriented translation solutions

As explained in Section 3, the use of passive is a register feature of English popular-scientific texts, whereas German prefers passive alternatives. Thus, for the adequate translation of such texts, it is important to know how the register features are typically realised in the target language. Moreover, the possibility to view and learn from examples which illustrate cross-linguistic register differences would help the translator to find appropriate translation solutions. Such examples can easily be found in a database in which register-specific translation equivalents are stored. This database serves as reference guiding the translator through the consistent translation re-occurring register divergences. Figure 4 shows a query for the English register feature passive and the German corresponding constructions in the Translator's Workbench.



```
<segment features="unit passive agentless" comment="" ignore=0>40  But if the phenomenon was to be
searched for in other organisms,</segment>    <segment features="unit active" comment="" ignore=0> not all of
which like</segment>    <segment features="unit ed-nonfi-construction ed-nonfin-clause withoutagent"
comment="" ignore=0> being lined up in tubes,</segment>    <segment features="unit ing-nonfinite-construction
ing-nonfiniteclause" comment="" ignore=0> some method of resolving the electrical activity in space of single
cells</segment>    <segment features="unit passive agentless" comment="" ignore=0> had to be
devised.</segment>

<segment features="unit passivariation man" comment="" ignore=0>40  Wollte man diese Erscheinung auch an
anderen Organismen nachweisen,</segment>    <segment features="unit passivariation lassen+sich+infinitiv
noagens" comment="" ignore=0> von denen sich nicht alle in Roehren aneinanderreihen lassen,</segment>
<segment features="unit passif ohneagens" comment="" ignore=0> musste eine Methode gefunden
werden,</segment>    <segment features="unit passivariation lassen+sich+infinitiv noagens" comment=""
ignore=0> mit der sich Elektrizitaet im Umfeld einzelner Zellen aufnehmen liess.</segment>
```

Figure 4: Translation of register features

As can be seen in Figure 4, the English register feature passive is translated into passive alternatives, which are characteristic for the

German register of popular-scientific texts (emphasised in circles). Furthermore, there is also an English active construction which has been translated into a German passive alternative (emphasised in a rectangle). This example clearly shows how the use of a parallel corpus annotated with register features can help the translator to solve register-specific translation problems and thus to produce a target language text adequate for the given register in the given language.

Summary and conclusion

The aim of this paper was to show how translation corpora annotated with register features can be used to solve translation problems. We tried to emphasise the importance of this topic for the translator's daily life, explained the relevance of cross-linguistic register differences, introduced the necessary computational techniques and discussed a sample solution for the translation of register-specific translation problems.

However, there are some remaining problems: most of the annotation and concordance tools do not operate on parallel corpora; and if they do so, they only allow the extraction of words and/or part-of-speech tags. The extraction of more abstract linguistic units with translation memories is not user-friendly since the tags are always displayed and the matching strings can hardly be identified. Another problem is that queries on source language words/tags and target language words/tags at the same time are not supported yet. Furthermore, in most of the cases the input and output formats of the different tools do not match and have to be transformed in order to be re-usable for other tools.

A possible solution to almost all of these problems is the use of an integrated XML version of the annotated corpus (cf. Teich & Hansen 2001). This allows the inclusion of alignment and meta-information as well as the integration of several layers of annotation into the corpus. Furthermore, it is possible to validate the annotated XML corpus against its DTD (Document Type Definition), i.e. a formal annotation grammar. Finally, XSLT or other XML-based query languages can be used for the extraction of every kind of information contained in the corpus as well as the querying of complex combinations of corpus contents and annotations.

NOTAS

[1] http://www.atril.com

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

REFERÊNCIAS BIBLIOGRÁFICAS

BIBER, D. *Dimensions of register variation.* Cambridge: University Press, 1995.

CHRIST, O. A modular and flexible architecture for an integrated corpus query system. *Proceedings of COMPLEX 94. 3rd Conference on Computational Lexicography and Text Research*. Budapest, 1994. p. 23-32.

FLUCK, H.-R. *Fachdeutsch in Naturwissenschaft und Technik*. Heidelberg: Julius Groos, 1997.

O'DONNELL, M. From Corpus to Codings: Semi-Automating the Acquisition of Linguistic Features. In: *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*. Stanford: Stanford University, 1995. p. 120-124.

SAGER, J. C., DUNGWORTH, D, MCDONALD, P. F. *English Special Languages.* Wiesbaden: Oscar Brandstetter, 1980.

SCHWANZER, V. Syntaktisch-stilistische Universalia in den wissenschaftlichen Fachsprachen. In: BUNGARTEN, T. (Ed.). *Wissenschaftssprache.* München: Wilhelm Fink, 1981. p. 213-230.

STEINER, E. Translation English – German: investigating the relative importance of systemic contrasts and of the text-type „translation". *SPRIK-Reports*, Oslo, No. 7, p. 1-48, 2001.

TEICH, E. *Contrast and commonality between English and German in system and text.* Saarbrücken: Philosophische Fakultät II, Universität des Saarlandes, 2001. (postdoctoral thesis)

TEICH, E., HANSEN, S. 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, 2001.