Multi-layer analysis of translation corpora: methodological issues and practical implications

Silvia Hansen
Department of Applied Linguistics,
Translation and Interpreting
Universität des Saarlandes
66041 Saarbrücken, Germany
`S.Hansen@mx.uni-saarland.de`

Elke Teich
Department of Applied Linguistics,
Translation and Interpreting
Universität des Saarlandes
44041 Saarbrücken, Germany
&
Department of Linguistics
University of Sydney, Australia
`E.Teich@mx.uni-saarland.de`

Abstract

The present paper discusses an application of multilingual, multi-layer corpus analysis from translation studies. The concrete context is the empirical testing of hypotheses about the specific properties of translations, such as explicitation, simplification, sanitization or normalization. While some of these assumed properties can be tested using some rather shallow measures that operate at the level of words (e.g., type-token ratio or lexical density), others require the analysis of more abstract linguistic features at the level of the clause; also, it is necessary to refer to linguistic features across different strata of linguistic organization (semantics, grammar, discourse).

We present some selected hypotheses about the nature of translations which have been tested on a bilingual English-German corpus. We discuss the particular corpus design adopted and the requirements on the techniques needed to extract information from this corpus and show how we use a combination of some standard corpus analysis techniques, such as string-based concordancing and part-of-speech tagging, and semi-automatic corpus analysis tools. We conclude with a summary and a list of issues for future work.

1 Introduction

The present paper discusses an analysis scenario in which instances of features of various linguistic levels need to be extracted in more than one language. The concrete application discussed is one from translation studies. Our main theoretical goal is the empirical testing of hypotheses about the specific properties of translations (see below). As a secondary goal, we want to elaborate the specific requirements on tools and techniques for a corpus-based analysis of translations and ultimately build up a translation corpus workbench that caters for the specific informational needs of researchers, teachers and students in translation studies. While there is a lot of experience in the construction of monolingual corpus resources (*tree banks*; e.g., (Marcus et al, 1993) for English, or (Brants, 1999; Plaehn, 2000; König and Lezius, 2000) for German), linguistically-annotated multilingual corpora remain rare, and translation corpora even rarer (with the exception of corpora for statistical machine translation and translation memories (example-based translation), but these are typically "raw" or only shallow-annotated).

It is commonly assumed in translation studies that translations are specific kinds of texts that are not only different from their original source language (SL) texts, but also from comparable original texts in the same language as the target language (TL). For instance, it has often been observed that translations tend to be longer than their SL originals, on the one hand, and that they are simpler than their SL originals or than comparable original texts in the TL, on the other hand. There has recently been an increased interest in more exact formulations of such general contentions in terms of explicit hypotheses and in providing empirical evidence to confirm or reject them. Such formulations can be found for example in (Toury, 1995), (Baker, 1995) or (Kenny, 1998). In (Baker, 1995), for instance, the following hypotheses are formulated:

- **Simplification**. Translations tend to use simpler language than original texts in the same language as the TL, possibly to optimize the readability of the target language text. Possible measures for simplification are average sentence length, lexical density and type-token ratio, the latter being a standard measure for the vocabulary variation in a text.

- **Explicitation**. Translations show a tendency to spell things out rather than leave them implicit. A possible measure for explicitation is text length, translations tending to be longer than their SL originals or monolingually comparable original texts. Also, some language-specific tests have been proposed, e.g., for English, frequency counts of optional *that* (both as complementizer and as relative pronoun) have been suggested, translations tending to use *that* more frequently than comparable original texts.

- **Normalization**. Translations have a tendency to conform to the typical patterns of the TL, exaggerating the typical features of the TL. As a test for normalization, (Baker, 1995) suggests comparing the use of punctuation, translations purportedly using punctuation less creatively than comparable texts in the same language as the TL.

- **Levelling-out**. In a collection of translations compared to a collection of comparable original texts in the same language as the TL, the individual texts in the set of translations are more similar to each other than the individual texts in the set of original texts. For levelling out, the above mentioned measures can be applied: For translations one would predict that the extreme values for lexical density, type-token ratio and average sentence length are closer to each other than for original texts.

While Baker's proposal is clearly a novel idea of how to approach the question of the specific properties of translations, there are a number of shortcomings: First of all, the properties of translations are only analyzed in relation to monolingually comparable texts. This disregards one of the major features characterizing the process of translation, namely that it is a process of text-induced text production, where this text is in another language. Possible interferences between source and target language, which may also contribute to making translations a special kind of texts (cf. (Toury, 1995)'s 'law of interference') can thus not be considered. Second, the measures suggested for testing the hypotheses are quite shallow linguistic properties exclusively, essentially operating at word and graphological levels.

Our own approach to the question about the specific properties of translations (cf. (Hansen, 1999; Steiner, 2001; Teich, 2001)) differs from this in the following respects: First, we take into account the source language as well; second, we bring some more abstract linguistic features into the picture. This allows us to formulate additional questions about the properties of translation, such as

- Do translations compromise the contrastive-typological features of the source language and the target language and therefore appear to be "normalized" or "simplified" or "more explicit"?

- Do translations compromise the contrastive register features of source language and target language texts and therefore appear to be "normalized" or "simplified" or "more explicit"?

This re-orientation towards the contrastive properties of source language texts and their translations has the following implications for the methods of analysis: First, additional hypotheses about the specific nature of translations relating to the source language can be formulated. One such hypothesis is the following:

- In translations, the **source language tends to shine through**.

This implies that we need a corpus that includes source language texts as well as monolingually comparable texts. Second, the features selected for analysis are generally more abstract than type-token ratio or lexical density because they are taken from the pool of the contrastive-typological and contrastive-registerial features of the source language and the target language; and third, because one source from which we select linguistic features for hypothesis testing are register features, the corpus is registerially controlled.

In the remainder of the paper we go through four hypotheses concerning the specific properties of translations, showing how we need to refer to multiple levels of linguistic organization, in order to be able to extract from the corpus instances of linguistic features ranging from grammatical (both structural and functional) to semantic ones (Section 2.1). We then show which kinds of techniques we have employed for dealing with which kind of analysis task and how these techniques combine in a more or less straightforward way (Section 2.2). Section 2.3 briefly assesses the techniques employed and discusses some requirements on a corpus analysis framework for contrastive-linguistic analysis, including the analysis of source language texts and their translations. Section 3 concludes the paper with a summary and issues for future work.

2 A sample analysis of some specific properties of translations

The specific properties of translations we have selected for presentation here are simplification, explicitation, normalization and shining-through. In the following we provide more precise formulations of these hypotheses relating to German translations from English and give at least one possible test for each. The tests illustrate the extraction of instances of linguistic features on the basis of a raw text corpus, on the one hand, and a multiply annotated corpus, on the other hand. For each test, both the annotation method (automatic vs. manual; different linguistic levels of annotation) and the extraction technique is described.

The concrete corpus design we adopt takes account of the requirements formulated above (Section 1): The corpus consists of German translations (from English) and comparable original German texts, *as well as* the English original source language texts of the German translations, and it is registerially controlled, i.e., it is partitioned into several registers (the registers in focus here are prepared political speeches and popular-scientific texts). The corpus is currently being encoded with metainformation (bibliographical data and registerial information, e.g., information about the domain, as well as document structure information) using XML and following the TEI standards (cf. (Teich and Hansen, 2001)). For the present analysis, each subcorpus (English originals, their translations into German, German comparable original texts) is represented by 10,000 words of texts, which follows a recommendation given in (Biber, 1995), who maintains that it is possible to represent the distribution of most linguistic features of a particular register on the basis of relatively short text samples (1,000 words) and relatively few texts from the given register (10 texts) (Biber, 1995, 131).[1] The corpus design is graphically displayed in Figure 1. On this basis, the following relations can be investigated: the relation between translations and original texts the same language, i.e., *monolingually comparable texts* (G-T—G-O), the relation between SL texts and their translations, i.e., *parallel texts* (E-O—G-T), and the relation between original texts in more than one language, i.e., *multilingually comparable texts* (E-O—G-O). Here, we focus on the analysis of the monolingually comparable corpus, but the analyses of the parallel and the multilingually comparable corpora are referred to in the interpretation of the analysis results.

2.1 Formulation of hypotheses

Simplification. Simplification means that translations tend to use simpler language than comparable original texts (cf. Section 1). Taking lexical density as a possible measure providing evidence for simplification, the following more con-

[1] An experiment made in (Biber, 1990) shows a reliability coefficient of > 0.80 for 1,000 word texts extracted from larger texts of several registers, and reliability coefficients of > 0.90 and > 0.95 for five-text samples and ten-text samples, respectively.

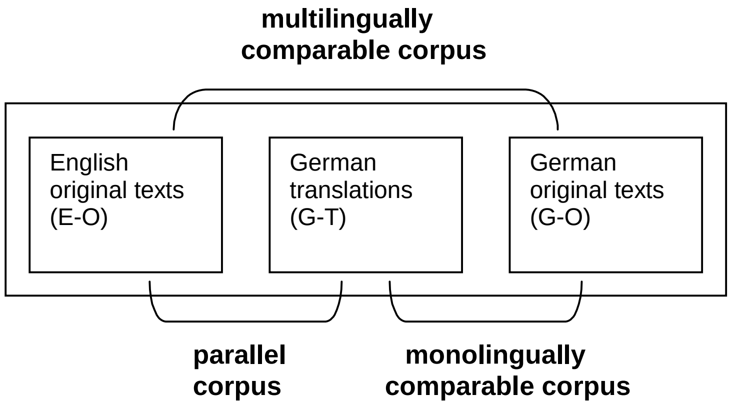

multilingually comparable corpus
English original texts (E-O)
German translations (G-T)
German original texts (G-O)
parallel corpus
monolingually comparable corpus

Figure 1: Corpus design

crete hypothesis can be formulated (cf. Baker95):

> (H1) In translations from English into German, one would expect the lexical density to be lower than in original German texts.

Explicitation. Translations tend to be more explicit than comparable original texts in the same language as the TL. For instance, they may use more optional elements (cf. (Baker, 1996, 180)), such as *daß/that* as complementizer, or employ more explicit (less densely packed) linguistic renderings of a given semantic content vs. less explicit ones (more densely packed), e.g., more conjunctions vs. prepositions. Conjunctions indicate that logico-semantic relations, such as temporal or causal ones, are made explicit, prepositions indicate a less explicit lexico-grammatical rendering of such relations (cf. (Halliday and Matthiessen, 1999)). See examples (1) and (2) for illustration.[2]

(1) *After she had arrived at the airport, she caught a bus to Central Station.*

(2) *After arrival at the airport she caught a bus to Central Station.*

Taking the number of occurrences of *daß/that* complementizers and the number of occurrences of conjunctions vs. prepositions as indicators for explicitation, we can formulate the following hypotheses:

> (H2) In translations from English into German, one would expect to find more *daß* complementizers than in original German texts.

> (H3) In translations from English into German, one would expect to find more conjunctions and fewer prepositions than in original German texts.

Normalization/Shining-through. One of the register features of the register of scientific writing in English that also holds for popularized scientific texts is the extensive use of passive. In German, passive is a typical feature of scientific texts as well, but there is also a wide range of other constructions which fulfil a similar function (i.e., not specifying or underspecifiyng the Agent of a process). These constructions occur frequently in German original texts of the given register, e.g., constructions with *lassen* (*let*) plus a reflexive verb, impersonal constructions with *man* (*one*) or *sein zu* (*be to*) constructions; henceforth these will be called passive alternatives. For some examples see (3)-(5) below.[3]

(3) *Somit lassen sich auch bei*
thus let themselves also with
diesen Spielen verschiedene Strategien
these games different strategies

[2] Note that often, a more explicit grammatical realization of a given content implies a register down shift; e.g., example (2) would be considered appropriate in more formal contexts, whereas (1) is rather neutral.

[3] We provide interlinear English glosses as well as translations. The German examples are taken from the G-O corpus; the English translations are taken from a corpus of translations of the G-O texts.

in einer Auszahlungsmatrix einander
in one pay-off-matrix to-another
gegenüberstellen und bewerten.
oppose and evaluate

"For these games, too, it would be possible to set up a pay-off matrix for comparing and evaluating different strategies."

(4) *Man hat dann ausserdem für eine*
one has then in-addition for an
Soforthilfe noch immer das
immediate-aid still always the
Strophantin zur Verfügung.
strophanthin at disposal

"If however a rapid, preferably instant, effect is desired, strophanthin will be needed."

(5) *Dabei ist eine sehr bemerkenswerte*
here is a very remarkable
Verlagerung der Schwerpunkte zu
shift in focus to
verzeichnen.
note

"There has also been a remarkable shift in emphasis."

Taking the extensive use of passive to be 'normal' in English in the given register, we can formulate the following hypothesis:

> (H4) In translations from English into German, one would expect that if English shines through, there will be a higher frequency of passive in German translations than in German comparable texts; if the German translations are normalized, the frequency of passive alternatives will be higher than in German originals.

Other features giving an indication of what is 'normal' in a given language are typological features. A commonly acknowledged contrastive-typological feature of the English and German grammatical systems is that English is both transitively and ergatively organized (Halliday, 1985), the former encoding the involvement of an Agent in a process, the latter encoding the lack of an Agent. English shifts between the two kinds of organization, whereas German is more strictly transitively organized—the middle variant is much rarer and more complex morpho-syntactic means are required for realization (notably reflexivization); cf. (Hawkins, 1986)). For illustration see examples (6)-(7) below.

(6) *He sold the book. — The book sold well.*

(7) *Er verkaufte das Buch. — Das Buch*
he sold the book — the book
verkaufte sich gut.
sold itself well

"He sold the book." — "The book sold well."

One would thus expect that middles occur more frequently in English texts than in German texts. On this basis we can formulate another hypothesis concerning normalization/shining-through:

> (H5) In translations from English into German, one would expect that if English shines through, there will be a higher number of middles in German translations than in German comparable texts; if the German translations are normalized, the number of middles will be lower in German translations than in German originals.

For the empirical testing of hypotheses (H1)–(H5), we need to refer to a range of more or less abstract linguistic features: Lexical density operates with the notion of function words vs. lexical words, *daß/that* is simply a string, conjunctions and prepositions are parts of speech (PoS), passive is a syntactic construction and the pair transitive-middle refers to two functional-grammatical classes that reflect a semantic distinction. In the following section we describe how we analyzed the corpus so that we can extract the information referred to in hypotheses (H1)–(H5).

2.2 Techniques of analysis and analysis results

The empirical testing of each of the hypotheses formulated above places different requirements on the corpus analysis techniques to be used.

	E-O	G-T	G-O
lexical density	50.62	48.67	49.55

Figure 2: Lexical density in E-O, G-T and G-O (political speeches)

While calculating lexical density and carrying out string searches are standard functions of a concordance tool, such as WordSmith (Scott, 1996), and operate on raw texts, for the others, various kinds of annotation are needed to extract the desired information, and these annotations can only be partly carried out automatically.

Lexical density. For determining lexical density, we use the WordList function of the WordSmith Tools, which matches a function word list (here: for English and German) with the words in a corpus. On this basis, lexical density can be calculated. The results for lexical density for the register of political speeches for all of the English originals, the German translations and the German comparable texts are shown in Figure 2. According to (H1), the lexical density in translations should be lower than in a monolingually comparable corpus. The analysis shows that in fact, lexical density is lower in the German translations, if only slightly, and we can thus interpret this result as simplification. In addition, this result may be interpreted as normalization and count as counter evidence to shining-through.

Optional daß/that. For the analysis of explicitation on the basis of optional vocabulary, such as the optional complementizer *daß/that*, we use the Concord function of WordSmith for the monolingual analysis and ParaConc (Barlow, 1994) for the bilingual analysis. To count occurrences and omissions of *daß/that*, we have created concordances of the most frequently used public verbs (e.g., *say/sagen*, *tell/sagen*, *declare/erklären* etc.). The results are displayed in Figure 3.[4] According to (H2), the translations show explicitation using the optional *daß* more often than comparable texts do. The analysis results show the contrary tendency of explicitation: There are significantly fewer *daß*-clauses in G-T compared to G-O and more deletions of *daß*. However, compared to their source language texts, the German translations contain more *daß*-clauses and fewer *daß*-deletions, so there is explicitation in relation to the SL texts. The use of the optional complementizer *daß/that* is an indicator of shining-through as well because a typical pattern of the English originals (i.e., in this case the frequent use of *daß/that*-deletion) is reproduced in the German translations.

	E-O	G-T	G-O
that/daß	3	5	13
that/daß deletion	19	12	8

Figure 3: Optional *that/daß* in E-O, G-T and G-O (political speeches)

Conjunctions vs. prepositions. In order to extract instances of a particular word class, a corpus needs to be annotated with parts of speech. The tagger we employ for PoS-tagging is the TnT tagger (Brants, 2000a). TnT has been chosen over other possible taggers because it has a good reliability (about 97 percent) and it has been trained for both German and English. The tag sets used are the Susanne tag set for English (Sampson, 1995) and the Stuttgart-Tübingen tag set for German (Hinrichs et al, 1995).

For the extraction of text instances annotated with part of speech information we employ the IMS Corpus Workbench (Christ, 1994). Importing TnT output to the workbench is a straightforward step. The required information can then be extracted using the workbench's Corpus Query Processor (CQP), which implements a query language on the basis of regular expressions. The results are displayed in Figure 4. Hypothesis (H3), which says that there should be more conjunctions in the translations for explicitation to hold, is confiremd, because the ratio of conjunctions to prepositions is 1 : 4.9 in G-T and 1 : 5.1 in G-O.

Passive. For the extraction of instances of particular syntactic constructions, we also employ the IMS Corpus Workbench. Typically, several different queries need to be applied to get all (satisfactory recall) and only the relevant matches (satisfactory precision), and in some cases irrel-

[4] Note that all frequency counts have been normalized on the basis of 10,000 words. Also, a significance test (chi-square) has been applied to all of the frequency counts except for lexical density. For *daß/that*, passive and transitive/middle all results are significant (cf. (Teich, 2001)).

	E-O	G-T	G-O
conjunctions	142	216	182
prepositions	1364	1078	936

Figure 4: Number of conjunctions in E-O, G-T and G-O (popular-scientific texts)

	E-O	G-T	G-O
passive	165	100	79
active	278	357	389
passive alternatives	64	180	146

Figure 5: Active, passive and passive alternaties in E-O, G-T and G-O (popular-scientific texts)

evant matches have to be removed from the list manually. Also, the queries need to be different for English and German because (a) the tag sets are different and (b) English and German have different word orders. For a sample query for English passive and a corresponding concordance list see Figure 6.

Figure 5 shows the results of the analysis of passives. There is a significant difference in the use of passive and passive alternatives in E-O compared to G-O, English using more passives than German, so the extensive use of passive is more a register feature of English than of German, which uses passive alternatives more extensively than English in this register. Comparing G-O and G-T, we find, however, that G-T has a higher frequency of passive than G-O—thus, we encounter a case of source language shining-through in the German translations. At the same time, there is normalization concerning the use of passive alternatives because they are used significantly more frequently in G-T compared to G-O.

Middle/transitive. A feature such as middle/transitive is a functional-grammatical feature that encodes a process as being caused by an Agent (transitive) or not (middle) (cf. Section 2.1). It is hardly possible to extract instances of transitive and middle processes from a text that is only annotated with PoS tags. Even a shallow parsing would not be sufficient. Possibly, extraction could be facilitated with a functional-grammatically parsed corpus, but there is no way of annotating a text in this way with a fully automatic procedure that is at the same time reliable enough. For dealing with functional information, such as the encoding of agency in the transitive/middle alternation, we have chosen to annotate the corpus manually using Coder (O'Donnell, 1995). Coder is a tool that supports the annotation of texts in terms of features that are organized as system networks (Halliday, 1985). The functionalities of the tool include support for the definition of coding schemes and for the annotation of texts with a defined scheme (see Figure 7 showing the scheme definition GUI) as well as a concordance function and some simple descriptive statistics. The coding record is kept in an XML/SGML-like format (which can potentially be mapped onto proper XML).

Figure 8 shows the results for the transitive/middle analysis. There are no significant differences between E-O and G-O concerning the use of transitive vs. middle. However, there is a significant difference between G-T and G-O, G-T having a higher frequency of middles than G-O. Because there is no difference between the originals (E-O and G-O), we cannot attest normalization or shining-through. The differences must be due to some other factor that influences the translations (cf. (Teich, 2001) for possible explanations).

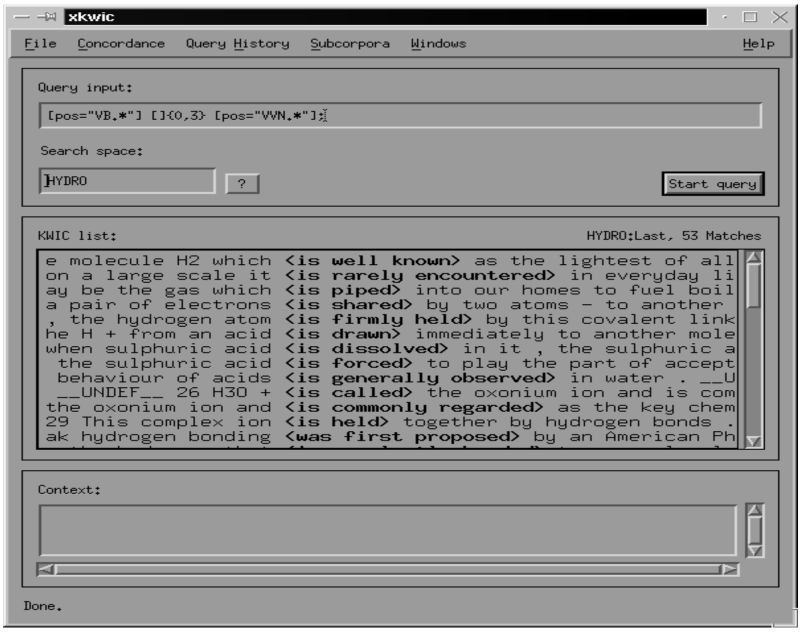
	E-O	G-T	G-O
transitive	504	510	481
middle	67	124	67

Figure 8: Transitive vs. middle in E-O, G-T and G-O

2.3 Summary and discussion

In Section 2.2 we have illustrated the analysis of a multilingual corpus for the purpose of testing some selected hypotheses about the specific properties of translations using some standard corpus analysis techniques. The application shows a number of limitations of the adopted methods of analysis as well as of the tools employed.

While word/string-based techniques are easy to carry out automatically, word-based analysis is a rather limited method of multilingual analysis (see the testing of lexical density or the testing



Query input:

```
[pos="VB.*"] []{0,3} [pos="VVN.*"];
```

Search space: HYDRO

KWIC list: HYDRO:Last, 53 Matches

```
e molecule H2 which <is well known> as the lightest of all
on a large scale it <is rarely encountered> in everyday li
ay be the gas which <is piped> into our homes to fuel boil
a pair of electrons <is shared> by two atoms - to another
, the hydrogen atom <is firmly held> by this covalent link
he H + from an acid <is drawn> immediately to another mole
when sulphuric acid <is dissolved> in it , the sulphuric a
 the sulphuric acid <is forced> to play the part of accept
 behaviour of acids <is generally observed> in water . __U
 __UNDEF__ 26 H3O + <is called> the oxonium ion and is com
the oxonium ion and <is commonly regarded> as the key chem
29 This complex ion <is held> together by hydrogen bonds .
ak hydrogen bonding <was first proposed> by an American Ph
```

Context:

Done.

Figure 6: Passive query and concordance in the IMS Corpus Workbench

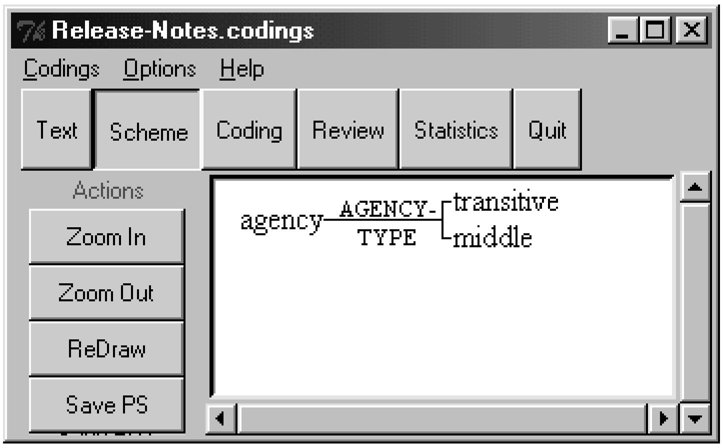


Figure 7: Coder: Annotation scheme definition

of optional *daß/that*). To extract instances of particular word classes (see the example of extraction of instances of conjunctions and prepositions in the previous section) or syntactic constructions (see the example of passive extraction above), the corpus needs to be at least annotated with part of speech information. While this is technically feasible, it needs to be noted as a disadvantage that extracting instances of syntactic constructions on the basis of PoS tags requires a lot of human effort in query formulation and possibly postediting. When it comes to more abstract features (see the example of transitive/middle extraction above), a PoS or shallow syntactically annotated corpus is not sufficient any more and the corpus has to be manually annotated. While there are computational tools supporting such annotations, manual analysis is still quite time-consuming and one needs to ensure inter-coder consistency.

The drawbacks mentioned so far are not as serious as to impede the feasiblity of the kinds of analysis illustrated in Section 2.2. However, there are a number of more serious problems to do with the fact that (a) we will need to make reference to more than one level of linguistic annotation at the same time, and (b) we deal with multilingual data that includes translations. Problems to do with multilingual data are first of all of a methodological kind, but they have implications for the techniques to be employed in corpus analysis and emerge at all levels of analysis, ranging from word counts on raw text to PoS-annotated and syntactically/semantically annotated text. First, comparing word counts cross-linguistically can be problematic when the languages involved exhibit substantial differences in their morphological typologies. In our case, English is analytic and German is highly inflecting. Word counts will therefore inherently reflect this typological difference. Second, with PoS-tagging, the tag sets employed must obviously be language-specific; but also, they should be similarly fine-grained in order to be cross-linguistically comparable. Similar constraints hold for shallow parsing. When explicit comparisons of translation units are to be carried out, parallel concordancing is needed. However, the alignment techniques publicly available (e.g., Déjà Vu (Atril, 2000) or Vanilla (Danielsson and Ridings, 2000)) do not yield entirely satisfactory results and human post-edition is required. When a parallel corpus is annotated, there is the additional problem of integrating the annotations with the alignment information. Among the tools discussed above, only the IMS corpus workbench allows integrating alignment with other kinds of annotation (here: PoS tags), so that parallel concordance queries are possible employing PoS tags.

The problem of integrating alignment with other kinds of corpus annotation is another facet of the more general problem of representation of multiply-annotated corpora. With the tools we have used for the analyses illustrated in Section 2.2, the corpus eventually exists in multiple versions, one for each kind of annotation in the worst case. It is thus not possible to pose queries to the corpus referring to more than one kind of annotation. If, for instance, we would like to verify the observation that English uses more nonagentive NP-Subjects than German (cf. (Hawkins, 1986; Doherty, 1993)) and formulate and test the hypothesis that in German translations this property of English may shine through, we would have to be able to refer to two different strata of linguistic organization, the semantic stratum (agentive/nonagentive) and the grammatical stratum (Subject, NP), and within the latter to a functional category and a surface-syntactic category. This is currently not possible.

Integrating different kinds of annotation into one representation has not only to do with different input/output formats, but also with the kinds of abstract internal representations adopted by the different tools and the different ways they encode information. For instance, Coder requires as input raw text and produces an SGML/XML-like representation of an annotated corpus, whereas TnT requires as input a tokenized text and represents a tagged text in a tab separated vector (TSV) format without further encoding. While part of this problem can be dealt with simply by format transformations (e.g., a TSV format can be straightforwardly transformed into an XML format by a Perl script, and an XML-like format can be straightforwardly transformed into XML, e.g., with the help of XSLT (W3C-XSLT, 2000)), there are some more principled questions involved here. If, for instance, clause annotations as we have done them using Coder are to be integrated with PoS annotations like the ones produced by TnT into one uniform representation, different units of annotation *and* different levels of annotation have to be integrated. Again, this would be feasible simply operating on the different formats, but in a more principled treatment, the units of annotation and their attributes would have to be defined explicitly in the first place. Similar problems arise in information extraction. Unless format transformations are carried out (where possible), different tools have to be employed and the corpus can only be queried with respect to one unit and one level of annotation at a time. It therefore seems to be desirable after all, to have a uniform representation that is built from first principles. This would require employing a document encoding standard in which an annotation grammar can be properly defined, as proposed for instance by XCES (XCES, 2000) or as implemented in the MATE system (Mengel, 1999; Mengel and Lezius, 2000).

3 Conclusions

The analysis task we are faced with in the corpus-based investigation of the specific linguistic properties of translations places a number of requirements on the methods and techniques to be used in corpus analysis (cf. Section 1). We have described the application of a set of computational techniques for the analysis of multilingual corpora both for linguistic annotation and information extraction (cf. Section 2.2). Taken together, the tools we have employed support the kinds of analysis we need to carry out, but there are a number of open issues that require more principled treatments.

The first issue relates to *corpus comparability*. Working on bilingual corpora that not only include translations into a TL, but also original texts in the TL, the question is how to make sure that the SL corpus and the corpus of TL original texts are comparable and that the corpus of TL translations and the corpus of TL original texts are comparable. For the present purposes we have applied a linguistic notion of comparability that is reasonably well defined, namely the notion of register. To ensure that two concrete corpora compiled on the basis of this notion are as comparable as possible, one could apply a statistical measure (as discussed e.g., in (Kilgarriff, to appear) who compares a number of statistical methods for measuring similarity between monolingual corpora).

The second issue has to do with enhancing the *reliability of the linguistic corpus analysis*: Since automatic annotation procedures are not entirely accurate, they would need to be systematically checked for errors. For instance, with a PoS tagger with a 97 percent reliability, there will be a tagging error in about every 20 words of texts. In a multilingual setting, error analyses would have to be done for all corpora and then compared across corpora. Possible negative effects on the reliability of extraction results of particular linguistic phenomena from the multilingual corpus could thus be anticipated and taken into account in the interpretation. Similar tests would need to be carried out for other kinds of automatic annotation, notably for shallow parsing, and also for manual annotation (cf. e.g., (Brants, 2000b)).

Third, apart from formulating and testing additional hypotheses about the specific properties of translations and testing them on a corpus basis, our longer-term goal is to work towards the specification of a *translation corpus workbench* that caters for the informational needs of researchers in translatology as well as teachers and students of translation. This involves testing other kinds of available techniques that promise to be suitable candidates for application to our particular analysis problems. On a slightly more abstract level, our goal is the integration of various kinds of corpus comparison, notably multilingual comparative analysis (of parallel corpora as well as of multilingually comparable corpora of original texts) and monolingual comparative analysis (here: of translations into a TL and original TL texts). A translation corpus workbench as we envisage it is thus not merely a ‘translation tree bank’ that contains a pair of monolingual tree banks with alignment—this is just one component. For instance, for concordancing, one of the desiderata when comparable original TL texts are taken into account in addition to SL texts and TL translations is to be able to search for the occurrence of a particular linguistic feature in three corpora in two languages at a time. Furthermore, in a corpus analysis setting such as the one described in this paper, multi-layer analysis is definitely required and would need to be supported by a translation corpus workbench. It seems to us that the issues that arise in the kind of multilingual analysis we have presented here are highly relevant for multi-layer corpus analysis more generally, both from a linguistic and from a computational point perspective. One of the linguistic issues involved here is the status of the notion of 'layer'. In one usage 'layer' refers to 'linguistic unit': word, phrase, clause, etc., where these units carry particular attributes (e.g., words have PoS as an attribute). Another usage of 'layer' refers to 'linguistic stratum': grammar, semantics, discourse; yet another refers to 'mode of discourse', i.e., spoken vs. written language, where spoken language involves the additional stratum of phonology and has its own units, such as the intonation unit or the syllable. From a linguistic point of view, these different notions would have to be kept conceptually apart. Multilingual analysis, like any other kind of comparative analysis, enforces such conceptual differentiation because the first thing one needs to do is to determine what it is exactly that is compared. Finally, from the point of view of information technology, the representation of multi-layer annotated multilingual corpus resources is presently one of the most challenging tasks in the area of integration of heterogeneous information sources in that it presents a highly interesting use case for emerging XML-based standards in document encoding as well as information extraction.

Acknowledgements

We are grateful for helpful comments by the anonymous reviewers.

References

Atril Development SL. 2000. Déjà Vu. Productivity system for translators. Software Manual. (`http://www.atril.com/`).

M. Baker. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–245, 1995.

M. Baker. 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*, pp. 175–186, Benjamins, Amsterdam.

M. Barlow. 1994. Paraconc. User manual. Technical report, Rice University, Houston, Texas.

D. Biber. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, 5:257–269, 1990.

D. Biber. 1995. *Dimensions of register variation: a cross-linguistic comparison*. Cambridge University Press, Cambridge.

T. Brants. 1999. Tagging and parsing with cascaded Markov models – automating corpus analyis. Technical report, Department of Computational Linguistics, Universität des Saarlandes, Saarbrücken. (`http://www.coli.uni-sb.de/ thorsten/tnt`).

T. Brants. 2000a. TnT —A statistical part-of-speech tagger. *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP) 2000*, Seattle, WA.

T. Brants. 2000b. Inter-annotator agreement for a German newspaper corpus. *Proceedings of the 2nd Conference on Language Resources and Evaluation LREC-2000*, Athens, Greece.

O. Christ. 1994. A modular and flexible architecture for an integrated corpus query system. *Proceedings of COMPLEX 94, 3rd Conference on Computational Lexicography and Text research*, Budapest, Hungary.

P. Danielsson and D. Ridings. 2000. Corpus and terminology: software for the translation program at Göteborgs Universitet or Getting students to do the work. In S. P. Botley, A. M. McEnery and A. Wilson, *Multilingual Corpora in Teaching and Research*, Rodopi, Amsterdam, pp. 65–72.

M. Doherty. 1993. Parametrisierte Perspektive. *Zeitschrift für Sprachwissenschaft*, 12(1):3–38, 1993.

M. A. K. Halliday and C. M. I. M. Matthiessen. 1999. *Construing experience through meaning. A language-based approach to cognition*. Cassell, London and New York.

M. A. K. Halliday. 1985. *An Introduction to Functional Grammar*. Edward Arnold, London.

S. Hansen. 1999. A contrastive analysis of multilingual corpora (English-German). Diploma Thesis, Department of Applied Linguistics, Translation and Interpreting, University of Saarland, Saarbrücken, Germany.

J. A. Hawkins. 1986. *A comparative typology of English and German*. Croom Helm, London and Sydney.

E. Hinrichs and H. Feldweg and M. Boyle-Hinrichs and R. Hauser. 1995. Abschlußbericht ELWIS. Korpusunterstützte Entwicklung lexikalischer Wissensbasen für die Computerlinguistik. Technical report, University of Tübingen, Germany.

D. Kenny. 1998. Creatures of Habit? What Translators Usually Do with Words. *Meta*, XLIII(4):515–524, 1998.

A. Kilgarriff. to appear. Comparing corpora. *International Journal of Corpus Linguistics*.

E. König and W. Lezius. 2000. A description language for syntactically annotated corpora. In *Proceedings of Coling 2000*, pp. 1056–1060, Saarbrücken, Germany.

G. P. Marcus and B. Santorini and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

A. Mengel and W. Lezius. 2000. An XML-based representation format for syntactically annotated corpora. *Proceedings of the 2nd Conference on Language Resources and Evaluation LREC-2000*, Athens, Greece.

A. Mengel 1999. Die integrierte Repräsentation linguistischer Daten. In J. Gippert (ed.), *Multilinguale Corpora. Codierung, Strukturierung und Analyse*. 11. Jahrestagung der GLDV, Prag, enigma corporation, pp. 115-121.

M. O'Donnell. 1995. From corpus to codings: semi-automating the acquisition of linguistic features. *Proceedings of the AAAI Symposium on Empirical Methods in Discourse Interpretation and Generation*, Stanford, California.

O. Plaehn. 2000. Computing the most probable parse for a discontinuous phrase structure grammar. Technical report, Department of Computational Linguistics, University of Saarland, Saarbrücken, Germany.

G. Sampson. 1995. *English for the Computer*. Oxford University Press, Oxford.

M. Scott. 1996. *WordSmith Tools Manual*. Oxford University Press, Oxford.

E. Steiner. 2001. Translations English – German: some observations on the relative importance of systemic contrasts and of the text type "translation". Paper presented at the Symposium on *Information Structure in a Cross-linguistic perspective*, University of Oslo, December 2000.

E. Teich and S. Hansen. 2001. Methods and techniques for a multi-level analysis of multilingual corpora. *Proceedings of Corpus Linguistics 2001*, Lancaster, UK.

E. Teich. 2001. Contrast and commonality in English and German system and text. A methodology for the investigation of the contrastive-linguistic properties of translations and multilingually comparable texts. Manuscript /submitted for publication), Department of Applied Linguistics, Translation and Interpreting, University of Saarland, Saarbrücken, Germany.

G. Toury. 1995. *Descriptive translation studies and beyond*. Benjamins, Amsterdam.

W3C. 2000. XSL Transformations (XLST). Version 1.0. (`http://www.w3c.org/TR/xslt`).

XCES. 2000. Corpus Encoding Standard for XML. 2000. Vassar College and LORIA/CNRS. (`http://www.cs.vassar.edu/XCES/`).