Linguistic enrichment and exploitation of the Translational English Corpus

Silvia Hansen-Schirra
Institute for Computational Linguistics
Saarland University, Germany
hansen@coli.uni-sb.de

1 Introduction

The aim of the present paper is the investigation of the nature of translated text. This kind of analysis is based on the assumption that translations differ from their source language texts and from comparable texts in the target language, in the sense that they have specific properties which cannot be found in non-translated text. The present research focuses on the comparison of English translations from the register of narrative writing to English originals from the same register. Within this context, the analysis of the relevant register features shows whether or not the translations conform to the norms of the given register, i.e. whether or not normalization can be found in the translation corpus (see section 2). The corpus under investigation is composed of the fiction part of the Translational English Corpus and the fiction part of the British National Corpus (see section 3). This monolingually comparable corpus comprises about 10 million words in total. To cope with the this large amount of words included in the comparable corpus, computational methods which support corpus enrichment and exploitation are employed. Along these lines, the automatic annotation of the corpus and its representational format are presented (see section 4). Furthermore, the querying techniques as well as the significance tests are discussed (see section 5). Hereby, the comparability of the two sub-corpora (i.e. translated vs. original text) plays an important role for the methodology chosen to investigate the specific properties of translations. The analysis results in a profile of the nature of translated text, which clearly shows to what extent the translations differ from originally produced texts. Furthermore, possible explanations are discussed on the basis of examples taken from the translation corpus. The paper concludes with a summary and an outlook on related research perspectives (see section 6).

2 The nature of translated text

In translation studies, the role of corpora was traditionally restricted to their use to the applied branch of this discipline. In particular, it has been used in the fields of terminology, translation aids (e.g., to develop translation memories or machine translation programs), translation criticism and translation training (to improve the final product with the help of corpus-based contrastive analysis and the study of *translationese*). Corpus linguistics has only very rarely been considered in terms of its importance to the theoretical and descriptive branches of translation studies. Researchers even tried to ban translations from corpora because translated text was regarded as inferior compared to originals and it was not considered worth investigating because it is generally constrained by the presence of a fully articulated text in another language. Sager (1984) was one of the first researchers who saw the need to examine translations as a special kind of text production and to look into their special characteristics. Nevertheless, he thought that the value of a translation is dependent on that of its original text (cf. Sager 1994). In contrast to Sager, Baker (1995) goes a step further by trying to exclude the influence of the source language on a translation in order to analyze characteristic patterns of translations independent of the source language. Within this context, Baker (1996) developed the following hypotheses on the universal features of English translations:

- **Explicitation.** Explicitation means that translators tend to render explicit implicit contents of the source language text in their translations. As a result, translated texts tend to contain less ambiguities than originally produced texts. Evidence of explicitation may, for example, be found in the text length (number of words of the individual texts), since, in many cases, translations are longer than texts produced originally in the target language or in the source language. This kind of analysis requires a comparison of source language texts and their translations on a text-by-text basis. Explicitation can also be analyzed in view of lexis and syntax, using a monolingual corpus of translated texts and a comparable corpus of original texts produced in the same language. Translations tend to use more explanatory vocabulary (e.g., ”therefore“, ”consequently“) and optional subordinators (e.g., ”that“) than in originally produced texts, thereby rendering implicit contents more explicit.

- **Simplification.** Simplification describes the tendency of translators to (consciously or unconsciously) simplify texts in order to improve the readability of their translations. Evidence of this tendency may, for instance, be found in the average sentence length: the mean sentence length of translated texts tends to be lower, as translators often break up long and complex sentences into two or more sentences in their translations in an effort to make the texts easier to read. Some linguistic features indicating simplification (e.g., the use of finite structures in English translations as opposed to non-finite structures in English originals) may also, at the same time, be a sign of explicitation. Another linguistic feature which reflects simplification is punctuation. Translators often change the punctuation from a weaker to a stronger mark, often using semicolons or periods instead of commas and periods instead of semicolons. This can be seen as an attempt to make texts easier and structure them more clearly by strengthening the punctuation. Another piece of evidence of simplification consists in the lexical density of a corpus, lexical density being the ratio of lexical vs. grammatical words. It is calculated by first subtracting the number of function words from the total number of words. The number of lexical words thus obtained is divided by the total number of words and then multiplied by 100. In translations, the lexical density tends to be lower than the lexical density of originals. This means that translations contain more function words and fewer lexical words than originals and are thus easier to read. A further method to test simplification is the type-token ratio, that is the ratio of different tokens vs. running words. This percentage is determined by dividing the number of lemmata (types) by the total number of words (tokens) and then multiplying the result by 100. Translators tend to use fewer types in translations than authors do in originals, and thus the type-token ratio of translations is lower than the type-token ratio of originals.

- **Normalization.** Normalization (or conservatism) means that translators tend to conform to the typical patterns of the target language or even to exaggerate their use. If, however, the status of the source language has an influence on the language use of the target language (like the influence of the English language on other languages in the area of software), normalization in translations is weakened, or even counteracted by a contrary tendency. If this is the case, the typical patterns of the source language are still visible in the translations. This universal feature also includes the tendency to normalize marked and ungrammatical structures. This often occurs in simultaneous or consecutive interpreting, where interpreters tend to finish unfinished sentences and to grammaticalize ungrammatical structures.

- **Levelling out.** In a corpus which consists of a sub-corpus of translations and a sub-corpus of texts originally produced in the target language, translations are more alike in terms of features such as lexical density, type-token ratio and average sentence length than the individual texts in the comparable corpus of source language and target language originals. This means that translators tend to use centered linguistic features, moving translations away from extremes.

Features like these translation universals constitute the specific properties of translations and thus nature of translated text.

3 Analysis scenario and corpus design

This section deals with the investigation of the relation between English translations and English originals, i.e. the investigation of translated text as special kind of text type or register. This analysis is based on Baker's normalization hypothesis (cf. Baker 1996), as the notion of register is used to substantiate the definition of what is ”normal“ (cf. Teich 2001). With this in view, Biber's lexico-grammatical register features (cf. Biber 1995) are taken into account. Since the corpus consists of English fiction texts, the following functional dimensions are of relevance: Dimension 2 (narrative vs. non-narrative discourse), Dimension 3 (situation-dependent vs. elaborated reference), Dimension 5 (abstract vs. non-abstract style) and Dimension 6 (on-line informational elaboration marking stance). Taking together all the sub-registers of fiction (general fiction, mystery fiction, science fiction, adventure fiction and romantic fiction), fiction can be characterized as narrative (i.e. the positive features of Dimension 2 are typical, while the negative ones are untypical), situation-dependent (i.e. the positive features of Dimension 3 are typical, whereas the negative ones are untypical), non-abstract (i.e. the positive features of Dimension 5 are typical, whereas the negative ones are untypical) and edited (i.e. the positive features of Dimension 6 are untypical, whereas the negative ones are typical). The combination of the functional dimensions relevant to fiction results in the following list of typical and untypical fiction features:

- typical features: past tense verbs, third person pronoun, perfect aspect, public verbs, synthetic negation, present participle clauses, time adverbials, place adverbials, adverbs, phrasal coordination;
- untypical features: present tense verbs, attributive adjectives, *wh*-relative clauses, pied piping, phrasal coordination, nominalizations, conjuncts, agentless passives, *by*-passives, past participle clauses, subordinators, *that*-clauses, demonstratives, final prepositions, existential *there*.

According to this list of lexico-grammatical features, a similar use (or overuse) of typical features as well as similar use (or underuse) of untypical features in English translated fiction texts (compared to English originals) would support Baker's normalization hypothesis since the target language texts would conform to (or even exaggerate) the norm of the target language. However, the notion of normalization has to be extended for this kind of analysis since the contrary tendency is investigated as well. The overuse of untypical features as well as the underuse of typical features in English translated fiction texts would therefore be seen as an indicator of *anti-normalization*, the contrary tendency of normalization, since the translated texts would not conform to the norms of the target language.

Since Baker's normalization hypothesis serves as basis for the investigation in this section, her criteria concerning corpus design are also taken into account (cf. Baker 1996). Thus, the corpus under investigation in this section consists of an English comparable corpus comprising of a sub-corpus of English translations and a sub-corpus of English originals. The sub-corpus of English translations is taken from the fiction part of the Translational English Corpus (TEC), whereas the sub-corpus of English originals is extracted from the fiction part of the British National Corpus (BNC). Thus, both sub-corpora belong to the register of fiction, as mentioned above. The translational sub-corpus is made up of translations from several languages into English (translated by professional translators who are English native speakers). The translational sub-corpus consists of 4,843,763 words and the original sub-corpus includes 4,741,500 words (9,585,263 words in total). From this it can be seen that the corpora are compiled in such a way as to make them as comparable as possible, both in terms of register and in terms of size.

4 Corpus enrichment

Since the corpus under investigation in this section is quite large (approximately 10 million words) and the features to be analyzed are rather shallow, linguistic annotation can be carried out automatically. Part-of-speech tagging is a fairly reliable method of annotation, either using a rule-based or a statistical approach. Recently, however, statistical approaches have become more popular. For this reason, the tagger which has been employed, the TnT tagger, is a statistical part-of-speech tagger that analyzes trigrams, incorporating several methods of smoothing and of handling unknown words (cf. Brants 1999). The system can be trained to deal with different languages and comes with the Susanne tagset for English (cf. Sampson 1995) and the Stuttgart-Tübingen tagset (STTS) for German (cf. Schiller et al. 1999). It includes a tool for tokenization, which is a preparatory step in the tagging process. In basic mode, not only does the tagger provide each token with a part-of-speech tag, but it omits alternative tags and also performs probability calculations. It analyzes between 30,000 and 60,000 tokens per second and has an accuracy of about 97 %.

A TEC sample output of TnT in tab separated vector (TSV) format is shown in figure 1. The part-of-speech tags used for tagging TEC and BNC are based on the Susanne tagset (cf. Sampson 1995).

He	PPHS1	and	CC
omitted	VVD	that	DD1
to	TO	,	YC
mention	VV0	as	CSA
that	CST	Romulus	NP1
one	MC1	had	VHD
person	NN1	quietly	RR
had	VHD	informed	VVN
been	VBN	him	PPHO1
killed	VVN	,	YC
and	CC	both	DB2
another	DD1	were	VBDR
badly	RR	past	II
wounded	VVN	saving	VVG
,	YC	.	VF

Figure 1: TnT sample output of TEC

There are two TEI-conformant headers for the files in TEC: a header for single volumes and a header for collected works. The header for single volumes can be found in the following figure:

```
<Header>
   <title>
      <filename></filename>
      <subcorpus></subcorpus>
      <collection></collection>
      <editor></editor>
   </title>
   <translator>
      <name></name>
      <gender></gender>
      <sexualOrientation></sexualOrientation>
      <Nationality></Nationality>
      <employment></employment>
      <status></status>
   </translator>
   <translation>
      <mode></mode>
      <extent></extent>
      <publisher></publisher>
      <pubPlace></pubPlace>
      <date></date>
      <copyright></copyright>
      <sponsor></sponsor>
      <reviews></reviews>
      <comments></comments>
   </translation>
   <translationProcess>
      <direction></direction>
      <mode></mode>
      <type></type>
   </translationProcess>
   <author>
      <name></name>
      <gender></gender>
      <sexualOrientation></sexualOrientation>
      <Nationality></Nationality>
   </author>
   <sourceText>
      <language></language>
      <mode></mode>
      <status></status>
      <publisher></publisher>
      <pubPlace></pubPlace>
      <date></date>
      <comments></comments>
   </sourceText>
</Header>
```

Figure 2: Header for single volumes in TEC

Each element in the headers has a start tag and an end tag. A start tag at the beginning of an element is represented by a balanced pair of angle brackets containing annotation strings, while a slash preceding the annotation strings indicates an end tag. The TEC header for single volumes includes the following information: title of the book, translator (where status refers to the question whether the translators have a full-time or part-time job and whether they work on a free-lance or in-house basis), translation (where extent means the number of words), translation process (where direction is relevant to whether or not the source language text is translated into the translator's mother tongue and type means a full translation in contrast to a summary, gist or excerpt), author of the original, source language text (where status refers to the problem whether the text is an original or a translation). The TEC header for collected works comprises an additional element called *section*. The section is repeated for each article or story contained in the collection. It includes information on the translator, the translation, the translation process, the author and the source language text of each article, story or paper in the collection.

In addition to the header, the body of each file contains meta-information on the text structure represented in a modified version of the Standard Generalized Mark-Up Language (SGML) (see figure 3 for the most important tags).

```
<title></title>
<head></head>
<sbhead></sbhead>
<chapter n="...">
<p>
<frontmatter></frontmatter>
<backmatter></backmatter>
<footnote></footnote>
<endnote></endnote>
<caption></caption>
```

Figure 3: Textual mark-up in TEC

For example, the following mark-up is included in each file of TEC: information on the (sub-)headings of the chapters or sections, the chapter number, the frontmatter (including introduction, preface etc.), the backmatter (including afterword, bibliography etc.), footnotes, endnotes, captions (e.g., for pictures or tables) and paragraphs[1].
The TEI-conformant header of the BNC includes the main elements displayed in figure 4:

```
<teiHeader>
  <fileDesc>
    <titleStmt></titleStmt>
    <editionStmt></editionStmt>
    <extent></extent>
    <publicationStmt></publicationStmt>
    <sourceDesc></sourceDesc>
  </fileDesc>
  <encodingDesc>
    <projectDesc></projectDesc>
    <samplingDecl></samplingDecl>
    <editorialDecl></editorialDecl>
    <tagsDecl></tagsDecl>
    <refsDecl></refsDecl>
    <classDecl></classDecl>
  </encodingDesc>
  <profileDesc>
    <creation></creation>
    <langUsage></langUsage>
    <particDesc></particDesc>
    <settingDesc></settingDesc>
    <textClass></textClass>
  </profileDesc>
  <revisionDesc>
    <change>
      <date></date>
      <respStmt></respStmt>
      <para></para>
    </change>
  </revisionDesc>
</teiHeader>
```

Figure 4: Header in BNC

The BNC header consists of a file description, an encoding description, a profile description and a revision description. The file description contains information on the title, the edition, the extent (i.e. size), the publication and the bibliographic source. The purpose of the coding project, the sampling criteria, the editorial principles, the linguistic annotation, the structure of the canonical references and the classification codes used for the texts within the corpus are spelled out in the encoding description. The profile description provides insight into the creation of the text, the language usage, the participants as well as their interaction, the settings of the communicative situation and the classification scheme with which the texts are categorized. The revision description explains the major changes which have taken place during the revision process.
The mark-up scheme of the BNC is an standardized SGML application (ISO 8879). The main elements of the textual mark-up describe the use of headings, segments, words, punctuation, texts, spoken texts and paragraphs. There are, however, other elements and attributes which describe the header and the markup of the BNC in greater detail[2].
In the BNC, all special characters are presented by SGML entity references, which take the form of an

[1] For further information on TEC refer to the following URL: http://ceylon.ccl.umist.ac.uk/.

[2] More information on the BNC can be found under the following URL: http://www.hcu.ox.ac.uk/BNC/.

ampersand followed by a mnemonic for the character and terminated by a semicolon (e.g., the representation *été* for the word *été*). In TEC, the transcription of these characters had to be carried out manually to make the data processable for annotation and querying tools. The use of international standards for the specification of application-independent document grammars (such as TEI for headers or SGML for mark-up) makes the corpora processable for computers as well as exchangeable and usable for researchers.

Since TEC had not been tagged before, the part-of-speech tags produced by TnT were added to the corpus. The BNC was tagged according to the CLAWS tagging scheme (cf. Garside 1987). However, in order to make the corpora as comparable as possible in terms of their linguistic interpretation, the CLAWS tags were removed and the BNC was tagged once again using TnT. For the representation of the part-of-speech tags in the comparable corpus, the vertical TSV output format was transformed into a horizontal format and added to the text, as can be seen from figure 5.

```
<body>
<w AT> The </w> <w NP1> Sheikh </w>
<w NN1> ' </w> <w FO> s </w> <w NN1> moustache </w>
<w VBDZ> was </w> <w ICS> like </w> <w AT1> a </w>
<w NN1> thistle </w> <w II> in </w> <w APPG> his </w>
<w NN1> bed </w> <w YC> , </w> <w AT> no </w>
<w NN1> matter </w> <w RRQ> how </w>
<w PPHS1> he </w> <w VVD> tossed </w>
<w CC> and </w> <w VVD> turned </w> <w YC> , </w>
<w PPHS1> he </w> <w RR> only </w>
<w VVD> rolled </w> <w AT> the </w> <w DAR> more </w>
<w II> on </w> <w II> to </w> <w PPH1> it </w>
<w YF> . </w> <w RR> Yet </w> <w PPHS1> he </w>
<w VBDZ> was </w> <w II> at </w> <w AT> the </w>
<w NN1> height </w> <w IO> of </w> <w APPG> his </w>
<w NN1> power </w> <w YF> . </w>
</body>
```

Figure 5: Part-of-speech tagging in TEC

In the horizontal tagging format, the part-of-speech tags are encoded as attributes of tokens.

5 Corpus exploitation

In order to find particular kinds of linguistic information in the corpus annotated in the ways described above, tools for querying the corpus for the features annotated are needed. For this purpose, the IMS Corpus Workbench (cf. Christ 1994) can be used. This concordance tool, with which it is possible to query for words and / or part-of-speech tags on the basis of regular expressions, consists of two modules: the corpus query processor (CQP) and the user interface (Xkwic).

Importing TnT output to the workbench is a straightforward step since the preparatory steps were all carried out by the part-of-speech tagger TnT (see section 4). These steps include character set normalization, tokenization and sentence boundary detection, such that the input format of the tagged corpus is TSV. The corpus is then encoded in such a way that it can be queried by the system. After the encoding of the corpus, all attributes (words, part-of-speech tags etc.) are declared in a registry file, which is a crucial element for all operations of the corpus maintenance.

The required information can then be queried using CQP, which implements a query language on the basis of the following regular expressions: concatenation, disjunction, negation, Kleene star, the plus and the interval operator. The results of the query are displayed using Xkwic. Figure 6 shows Xkwic with a query for passive. The query is based on the part-of-speech tags VB.* (forms of the verb *be*) followed by VVN.* (past participle) and zero to one words in between. The results are displayed in the KWIC (keyword in context) list indicating the number of matches. Xkwic offers the usual functionalities of a concordance program: concordances can, for example, be saved (all at once or in different sub-corpora), deleted, sorted and printed. The query history can be viewed, saved and loaded, and sub-corpora can be saved as well. Furthermore, an extended view on the KWIC concordances, a window for messages and an alignment window for parallel concordances can be displayed, too. Xkwic also calculates a frequency distribution for the first word or part-of-speech tag of the matches. The information thus provided makes it possible to analyze, for example, the use of aspect within passive constructions.

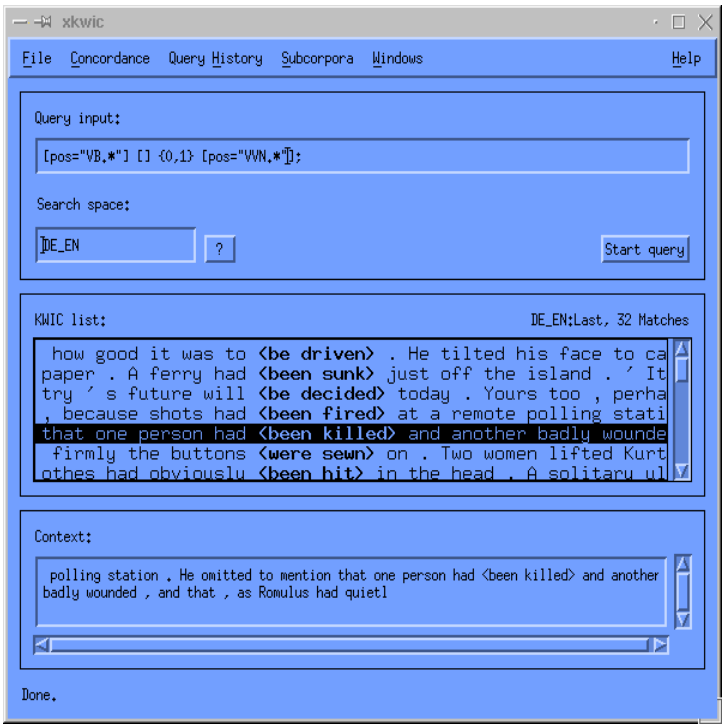

Figure 6: Passive query with Xkwic

As to the query of instances of complex syntactic constructions, typically several different queries need to be made to obtain all (satisfactory recall) and only the relevant matches (satisfactory precision). In some cases irrelevant matches have to be removed from the list manually. Thus, the most useful queries for the typical and untypical fiction features had to be found. There are three groups of queries: queries based on words, queries based on part-of-speech tags and queries based on both words and part-of speech tags which makes the queries quite complex.

Since the two comparable fiction parts in TEC and BNC analyzed in this paper are both in English, the queries can be used on both sub-corpora. This procedure ensures consistent querying even in cases where the values for precision and recall do not attain 100 %.

In order to attain comparable results, all frequencies of occurring features were normed using the following formula:

$$\frac{\text{frequency of feature occurrences x basis of norming}}{\text{word count}} = \text{normed frequency}$$

The basis of norming is 5 million for TEC and for BNC, this being the approximate size of the two sub-corpora.

As only the statistically significant results are relevant, all the frequencies of occurring features have to be subjected to a statistical test. The chi square test (cf. Oakes 1998) was employed in this paper. This test is a non-parametric statistical procedure for testing whether or not the distribution of feature occurrences is accidental. The first step of the test is to define the null hypothesis, according to which there is no statistically significant difference between the frequencies of feature occurrences found in the sub-corpora. If so, all the frequencies of occurrences would be the same as the sum of all frequencies divided by the number of categories. This value referred to as the expected value (E) is determined using the following formula:

$$\frac{\text{raw total x column total}}{\text{grand total of items}} = E$$

In contrast to the expected value, the actual frequency is called the observed value (O). With these two values the chi square value can be calculated as follows:

$$\Sigma\ (O\text{-}E)^2 : E = X^2$$

For the significance tests carried out in this paper, the two-tailed/non-directional chi square test was employed and, additionally, Yates's correction was calculated given that the amount of data under investigation is rather large. For the same reason, the level of significance was set to 0.001, which means that the results have to reach 99.9 % to be statistically significant.

6 The nature of TEC

The frequent use of typical register features is an indicator of normalization in translated texts. A comparison with English originals makes it possible to determine whether or not the use of fiction features in English translations obeys the usage norms. Of course the same holds true for the untypical fiction features, so that, if the English translations show instances of normalization, the untypical fiction features are expected to be underused. If the English translations do not obey the English usage norms, the typical features are underused, whereas the untypical features occur more frequently in the English translations than in the English originals.

In figure 7, the differences between the frequencies found in TEC and BNC are calculated and normed (as percentages). Thus, the positive percentages refer to the degree of normalization in connection with the typical features, whereas the negative percentages relate to the degree of anti-normalization for the typical features. The degrees of anti-normalization and normalization for the untypical features are summarized in figure 8.

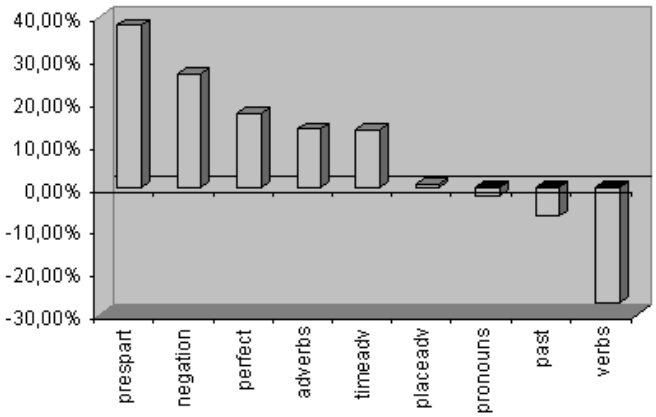

Figure 7: Degree of anti-/normalization for typical features

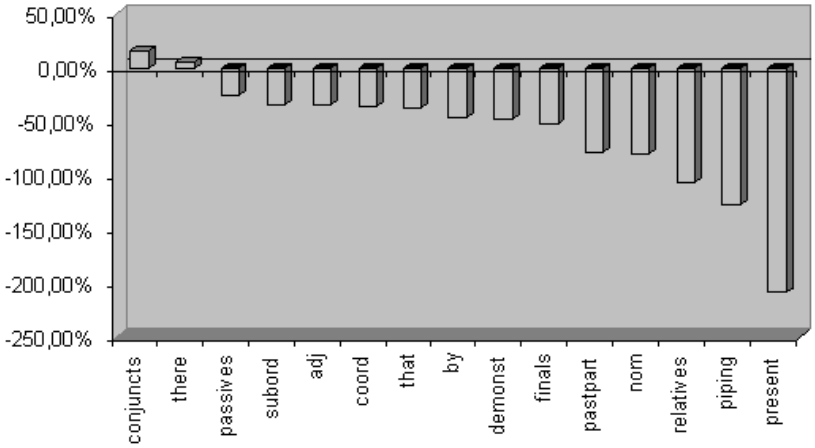

Figure 8: Degree of anti-/normalization for untypical features

As can be seen in figure 7, the higher frequency of typical fiction features to be found in TEC, as compared to the BNC, means that the translations show a trend towards normalization for typical fiction features. Untypical fiction features occur more frequently in TEC than in BNC, that is the translations show a trend towards anti-normalization for the untypical fiction features.

Thus, it can be observed that the typical fiction features are normalized and the untypical fiction features are anti-normalized in English translated fiction compared to English original fiction. This means that English translated fiction is more narrative and situation-dependent than English original fiction texts are (since an overuse of the typical features can be attested). At the same time, English translated fiction is more abstract and less edited than original fiction (since an overuse of the untypical features can be attested). This leads to the conclusion that translators tend to conform to the typical

patterns of the target language in terms of typical fiction features, the use thereof being exaggerated, whereas they tend to use the untypical fiction features less often. These observations show that a register shift between English original narrative texts and English translated narrative texts takes place in the sense that, owing to the extensive overuse of typical fiction features, English translated narrative texts are even more typical of their register. However, they point towards a more neutral register through the extensive overuse of untypical fiction features.

In the following, a sample concordance taken from TEC, the normalized use of the typical fiction feature *place adverbial* is exemplified:

- Marden is about to enter the house when he turns *around*: „You must leave the gun, naturally;
- there wasn't enough air to breathe. After he went *back* outside, on the road, he noticed that
- The bad weather will return. It was *there*, near the magazine (he can see it clearly from

Here, the *place adverbials* "around", "back" and "there" are instances of normalization, showing the overuse of this typical fiction feature. In the following, another sample concordance from TEC is displayed which illustrates the anti-normalized use of the untypical fiction feature *by-passive:*

- asked when it was all going to start, he *was pacified by* the chairman of the discussion:
- a black man's head horribly shrunken. Rebecca *was told all this and more by* Nell about a week after her arrival at Broom House.
- because I could hear voices in the corridor and I *was distracted by* the noise of the lift. It didn't

In this example, the *by-passives* "was pacified by", "was told all this and more by" and "was distracted by" contribute to anti-normalization, meaning that this untypical fiction feature is overused.

Reasons for normalization or anti-normalization regarding the typical and untypical fiction features are discussed in the following on the basis of examples from the German and French source language texts and their translations taken from TEC:

- German: Während *seiner Wanderung* durch den Wald dachte er an
- English: *he walked* through the forest and thought about

This example shows that the German nominalization "Wanderung" is translated into the English verb "walked". *Nominalization* being a feature untypical of English fiction, it was transferred into the typical fiction feature *past tense verbs*. In this case, register-specific language use causes an instance of normalization. A similar phenomenon can be found in the following example:

- German: Die Wahrheit über das dunkle Geheimnis fand man damals *in ihrem Tagebuch*.
- English: At that time, *her diary* explained the dark secret

Here, the typical fiction feature *past tense verb* is used both in the English and the German sentence, the semantic roles, however, are distributed differently as the German typology allows the frequent use of constructions like the impersonal passive alternative "man" (here: in combination with the PP "in ihrem Tagebuch"), whereas the English typology provides other lexico-grammatical means of expressing the same meaning. In this case, the non-agentive NP subject "her diary" is used. Therefore, typological differences are responsible the translational choices.

Those properties of translated text which are not register- or typology-related are inherent in the translation process itself. Evidence of this is contained in the following example:

- German: nicht mehr die einzige *Frau, die im Ort gefeiert wurde*.
- English: was no longer the only *celebrity in town*.

In this example, the German relative clause "die im Ort gefeiert wurde" is rendered into the English NP "celebrity in town". The English translation is less explicit that the source language equivalent because the fact that a woman is the only celebrity in town (and not a man) is implied by the context and not spelled out in the German text. This example clearly entails a loss of explicitation. In contrast to this, the following example illustrates the co-occurrence of simplification and explicitation:

- French: *Au début*, on crut à une bouderie d'enfant.
- English: *When they started*, it was thought to be a fit of childish

In this case, the French PP "Au début" is translated with the English subordinate clause "When they started". The use of the subordinate clause makes the English translation easier to read and more explicit, spelling out the meaning of the French PP.

The examples discussed in this section show that the different properties of translation intervene with each other. Furthermore, they can be in a causal relation towards each other with, for example, explicitation or simplification causing normalization.

7 Summary and outlook

The present paper described the linguistic enrichment and exploitation of the Translational English Corpus. This approach was based on Baker's prototypical hypotheses (section 2 and 3), but profited from linguistically enriched data (section 4). This means that lexico-grammatical features such as Biber's register features could be exploited automatically (section 5). The results of the investigation, i.e. the nature of the Translational English Corpus were presented in section 6.

Since normalization of the typical fiction features and anti-normalization of the untypical fiction

features could be found in English translated fiction, as compared to English original fiction, the following can be attested: English translated fiction is more narrative and situation-dependent than English original fiction texts (the typical features being overused). At the same time, English translated fiction is more abstract and less edited than original texts (the untypical features being overused). Thus, English translated narrative texts are more typical than English original narrative texts as regards the use of typical features, whereas they point towards a more neutral register owing to the extensive overuse of untypical fiction features.

On the basis of bilingual examples it could be shown that the sources of normalization and anti-normalization can partially be found in register- and typology-specific language use. The translation process, which triggers not only normalization and anti-normalization, but also other translation properties such as explicitation, simplification etc., can be seen as a further explanation.

The empirical analysis which was carried out in the framework of this paper has to be extended by taking into consideration the source language texts of the translations. This would allow the analysis of the influence of the source language on a broader basis and with empirical methods. Additionally, a sub-corpus of comparable texts in the target language could be added to the cross-linguistic approach. As a result, the translations could be compared to originals in the target language on the basis of a more profound linguistic analysis.

References

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

Baker M 1996 Corpus-based translation studies: The challenges that lie ahead. In Sommers H (ed), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam, John Benjamins, pp 175-186.

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

Brants T 1999 *TnT - A Statistical Part-of-Speech Tagger* (User manual). Department of Computational Linguistics, Universität des Saarlandes, Saarbrücken, Germany.

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

Garside R 1987 The CLAWS Word-tagging System. In Garside R, Leech G, Sampson G (eds), *The Computational Analysis of English: A Corpus-based Approach*. London, Longman.

Oakes M P 1998 *Statistics for Corpus Linguistics*. Edinburgh, Edinburgh University Press.

Sager J C 1984 Reflections on the didactic implications of an extended theory of translation. In Wilss W, Thome G (eds), *Translation theory and its implementation in the teaching of translating and interpreting*. Tübingen, Gunter Narr, pp 333-343.

Sager J C 1994 *Language engineering and translation: consequences of automation*. Amsterdam, John Benjamins.

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

Schiller A, Teufel S, Stöckert C 1999 *Guidelines für das Tagging deutscher Textcorpora mit STTS* (Technical report). Stuttgart and Tübingen, University of Stuttgart and University of Tübingen.

Teich E 2001 English-German *contrast and commonality in system and text. A methodology for the investigation of parallel and multilingually comparable texts* (Habilitationsschrift). Saarbrücken, Saarland University.