VARTRA: A Comparable Corpus for Analysis of Translation Variation

Ekaterina Lapshinova-Koltunski
Universität des Saarlandes
A 2.2 Universität Campus
66123 Saarbrücken
Germany
e.lapshinova@mx.uni-saarland.de

Abstract
This paper presents a comparable translation corpus created to investigate translation variation phenomena in terms of contrasts between languages, text types and translation methods (machine vs. computer-aided vs. human). These phenomena are reflected in linguistic features of translated texts belonging to different registers and produced with different translation methods. For their analysis, we combine methods derived from translation studies, language variation and machine translation, concentrating especially on textual and lexico-grammatical variation. To our knowledge, none of the existing corpora can provide comparable resources for a comprehensive analysis of variation across text types and translation methods. Therefore, the corpus resources created, as well as our analysis results will find application in different research areas, such as translation studies, machine translation, and others.

1 Introduction: Aims and Motivation
Comparable corpora serve as essential resources for numerous studies and applications in both linguistics (contrastive language, text analysis), translation studies and natural language processing (machine translation, computational lexicography, information extraction). Many comparable corpora are available and have been being created for different language pairs like (a) English, German and Italian (Baroni et al., 2009); (b) English, Norwegian, German and French (Johansson, 2002); (c) written or spoken English and German (Hansen et al., 2012) or (Lapshinova et al., 2012).

However, comparable corpora may be of the same language, as the feature of ‘comparability’ may relate not only to corpora of different languages but also to those of the same language. The main feature that makes them comparable is that they cover the same text type(s) in the same proportions, cf. for instance, (Laviosa, 1997) or (McEnery, 2003), and thus, can be used for a certain comparison task.

As our research goal is the analysis of translation variation, we need a corpus which allows us to compare translations, which differ in the source/target language, the type of the text translated (genre or register) and the method of translation (human with/without CAT tools, machine translation). There are a number of corpus-based studies dedicated to the analysis of variation phenomena, cf. (Teich, 2003; Steiner, 2004; Neumann, 2011) among others. However, all of them concentrate on the analysis of human translations only, comparing translated texts with non-translated ones. In some works on machine translation, the focus does lie on comparing different translation variants (human vs. machine), e.g. (White, 1994; Papineni et al., 2002; Babych and Hartley, 2004; Popović, 2011). However, they all serve the task of automatic machine translation (MT) systems evaluation and use the human-produced translations as references or training material only. None of them provide analysis of specific (linguistic) features of different text types translated with different translation methods.

The same tendencies are observed in the corpus resources available, as they are mostly built for certain research goals. Although there exists a number of translation corpora, none of them fits our research task: most of them include one translation method only: EUROPARL (Koehn, 2005) and JRC-Acquis (Steinberger et al., 2006) – translations produced by humans, or DARPA-94 (White, 1994) – machine-translated texts only.

1CAT = computer-aided translation
Moreover, they all contain one register only and, therefore, cannot be applied to a comprehensive analysis of variation phenomena.

Therefore, we decided to compile our own comparable corpus which contains translations from different languages, of different text types, produced with different translation methods (human vs. machine). Furthermore, both human and machine translations contain further varieties: they are produced by different translators (both professional and student), with or without CAT tools or by different MT systems.

This resource will be valuable not only for our research goals, or for research purposes of further translation researchers. It can also find further applications, e.g. in machine translation or CAT tool development, as well as translation quality assessment.

The remainder of the paper is structured as follows. Section 2 presents studies we adopt as theoretical background for the selection of features and requirements for corpus resources. In section 4, we describe the compilation and design of the comparable translation corpus at hand. In section 5, we demonstrate some examples of corpus application, and in section 6, we draw some conclusions and provide more ideas for corpus extension and its further application.

2 Theoretical Background and Resource Requirements

To design and annotate a corpus reflecting variation phenomena, we need to define (linguistic) features of translations under analysis. As sources for these features, we use studies on translation and translationese, those on language variation, as well as works on machine translation, for instance MT evaluation and MT quality assessment.

2.1 Translation analysis and translationese

As already mentioned in section 1 above, translation studies either analyse differences between original texts and translations, e.g. (House, 1997; Matthiessen, 2001; Teich, 2003; Hansen, 2003; Steiner, 2004), or concentrate on the properties of translated texts only, e.g. (Baker, 1995). However, it is important that most of them consider translations to have their own specific properties which distinguish them from the originals (both of the source and target language), and thus, establish specific language of translations – the translationese.

Baker (1995) excludes the influence of the source language on a translation altogether, analysing characteristic patterns of translations independent of the source language. Within this context, she proposed translation universals – hypotheses on the universal features of translations: explicitation (tendency to spell things out rather than leave them implicit), simplification (tendency to simplify the language used in translation), normalisation (a tendency to exaggerate features of the target language and to conform to its typical patterns) and levelling out (individual translated texts are alike), cf. (Baker, 1996). Additionally, translations can also have features of “shining through” defined by Teich (2003) – in this case we observe some typical features of the source language in the translation. The author analyses this phenomena comparing different linguistic features (e.g. passive and passive-like constructions) of originals and translations in English and German.

In some recent applications of translationese phenomena, e.g. those for cleaning parallel corpora obtained from the Web, or for the improvement of translation and language models in MT (Baroni and Bernardini, 2005; Kurokawa et al., 2009; Koppel and Ordan, 2011; Lembersky et al., 2012), authors succeeded to automatically identify these features with machine learning techniques.

We aim at employing the knowledge (features described) from these studies, as well as techniques applied to explore these features in the corpus.

2.2 Language variation

Features of translated texts, as well as those of their sources are influenced by the text types they belong to, see (Neumann, 2011). Therefore, we also refer to studies on language variation which focus on the analysis of variation across registers and genres, e.g. (Biber, 1995; Conrad and Biber, 2001; Halliday and Hasan, 1989; Matthiessen, 2006; Neumann, 2011) among others. Register is described as functional variation, see Quirk et al. (1985) and Biber et al. (1999). For example, language may vary according to the activity of the involved participants, production varieties (written vs. spoken) of a language or the relationship between speaker and addressee(s). These parameters correspond to the variables of
field, tenor and mode defined in the framework of Systemic Functional Linguistics (SFL), which describes language variation according to situational contexts, cf. e.g. Halliday and Hasan (1989), and Halliday (2004).

In SFL, these variables are associated with the corresponding lexico-grammatical features, e.g. field of discourse is realised in functional verb classes (e.g., activity, communication, etc) or term patterns, tenor is realised in modality (expressed e.g. by modal verbs) or stance expressions, mode is realised in information structure and textual cohesion (e.g. personal and demonstrative reference). Thus, differences between registers or text types can be identified through the analysis of occurrence of lexico-grammatical features in these registers, see Biber’s studies on linguistic variation, e.g. (Biber, 1988; Biber, 1995) or (Biber et al., 1999).

Steiner (2001) and Teich (2003) refer to registers as one of the influencing sources of the properties of translated text. Thus, we attempt to study variation in translation variants by analysing distributions of lexico-grammatical features in our corpus.

2.3 Machine translation

We also refer to studies on machine translation in our analysis, as we believe that translation variation phenomena should not be limited to those produced by humans. Although most studies comparing human and machine translation serve the task of automatic MT evaluation only, cf. (White, 1994; Papineni et al., 2002; Babych and Hartley, 2004), some of them do use linguistic features for their analysis.

For instance, Popović and Burchardt (2011) define linguistically influenced categories (inflections, word order, lexical choices) to automatically classify errors in the output of MT systems. Specia (2011) and Specia et al. (2011) also utilise linguistic features as indicators for quality estimation in MT. The authors emphasize that most MT studies ignored the MT system-independent features, i.e. those reflecting the properties of the translation and the original. The authors classify them into source complexity features (sentence and word length, type-token-ratio, etc.), target fluency features (e.g. translation sentence length or coherence of the target sentence) and adequacy features (e.g. absolute difference between the number of different phrase types in the source and target or difference between the depth of their syntactic trees, etc.).

3 Methodology

Consideration of the features described in the above mentioned frameworks will give us new insights on variation phenomena in translation. Thus, we collect these features and extract information on their distribution across translation variants of our corpus to evaluate them later with statistical methods.

Some of the features described by different frameworks overlap, e.g. type-token-ratio (TTR) or sentence length as indicator for simplification in translationese analysis and as a target fluency feature in MT quality estimation; modal meanings and theme-rheme distribution in register analysis and SFL, or alternation of passive verb constructions in register analysis and translation studies.

Investigating language variation in translation, we need to compare translations produced by different systems with those produced by humans (with/without the help of CATs). Furthermore, we need to compare translated texts either with their originals in the source or comparable originals in the target language. Moreover, as we know that text type has influence on both source and target text (Neumann, 2011), we need to compare different text registers of all translation types.

This requires a certain corpus design: we need a linguistically-annotated corpus for extraction of particular features (e.g. morpho-syntactic constructions); we need to include meta-information on (a) translation type (human vs. computer-aided vs. machine, both rule-based and statistical), (b) text production type (original vs. translation) and (c) text type (various registers and domains of discourse). This will enable the following analysis procedures: (1) automatic extraction, (2) statistical evaluation and (3) classification (clustering) of lexico-grammatical features.

4 Corpus Resources

4.1 Corpus data collection

Due to the lack of resources required for the analysis of translation variation, we have compiled our own translation corpus VARTRA (VARIation in TRANslation). In this paper, we present the first version of the corpus – VARTRA-SMALL, which is the small and normalised version used for our
first analyses and experiments. The compilation of the full version of VARTRA is a part of our future work, cf. section 6.

VARTRA-SMALL contains English original texts and variants of their translations (to each text) into German which were produced by: (1) human professionals (PT), (2) human student translators with the help of computer-aided translation tools (CAT), (3) rule-based MT systems (RBMT) and (4) statistical MT systems (SMT).

The English originals (EO), as well as the translations by professionals (PT) were exported from the already existing corpus CroCo mentioned in section 1 above. The CAT variant was produced by student assistants who used the CAT tool ACROSS in the translation process. The current RBMT variant was translated with SYSTRAN, although we plan to expand it with a LINGUATEC-generated version. For SMT, we have compiled two versions – the one produced with Google Translate and the other one with a Moses system.

Each translation variant is saved as a subcorpus and covers seven registers of written language: political essays (ESSAY), fictional texts (FICT), manuals (INSTR), popular-scientific articles (POPSCI), letters of shareholders (SHARE), prepared political speeches (SPEECH), and touristic leaflets (TOU), presented in Table 1. The total number of tokens in VARTRA-SMALL comprises 795,460 tokens (the full version of VARTRA will comprise at least ca. 1,7 Mio words).

### 4.2 Corpus annotation

For the extraction of certain feature types, e.g. modal verbs, passive and active verb constructions, Theme types, textual cohesion, etc. our corpus should be linguistically annotated. All subcorpora of VARTRA-SMALL are tokenised, lemmatised, tagged with part-of-speech information, segmented into syntactic chunks and sentences. The annotations were obtained with Tree Tagger.

In Table 2, we outline the absolute numbers for different annotation levels per subcorpus (translation variant) in VARTRA-SMALL.

VARTRA-SMALL is encoded in CWB and can be queried with the help of Corpus Query Processor (CQP) (Evert, 2005). We also encode a part of the meta-data, such as information on register, as well as translation method, tools used and the source language. A sample output encoded in CQP format that is subsequently used for corpus query is shown in Figure 1.

In this way, we have compiled a corpus of different translation variants, which are comparable, as they contain translations of the same texts produced with different methods and tools. Thus, this comparable corpus allows for analysis of contrasts in terms of (a) text typology (e.g. fiction vs. popular-scientific articles); (b) text production types (originals vs. translations) and (c) translation types (human vs. machine and their subtypes).

Furthermore, examination of some translation phenomena requires parallel components – alignment between originals and translations. At the moment, alignment on the sentence level (exported from CroCo) is available for the EO and PT subcorpora. We do not provide any alignment for further translation variants at the moment, although we plan to align all of them with the originals on word and sentence level.

### 4.3 Corpus querying

As already mentioned in 4.2, VARTRA-SMALL can be queried with CQP, which allows definition of language patterns in form of regular expressions based on string, part-of-speech and chunk tags, as well as further constraints. In Table 3, we illustrate an example of a query which is built to extract cases of processual finite passive verb constructions in German: lines 1 - 5 are used for passive from a *Verbzweit* sentence (construction in German where the finite verb occupies the position after the subject), and lines 6 - 10 are used for *Verbletzt* constructions (where the finite verb occupies the final position in the sentence). In this example, we make use of part-of-speech (lines 3a, 5, 8 and 9a), lemma (lines 3b and 9b) and

<table>
<thead>
<tr>
<th>subc</th>
<th>token</th>
<th>lemma</th>
<th>chunk</th>
<th>sent</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>132609</td>
<td>9137</td>
<td>55319</td>
<td>6525</td>
</tr>
<tr>
<td>CAT</td>
<td>139825</td>
<td>10448</td>
<td>58669</td>
<td>6852</td>
</tr>
<tr>
<td>RBMT</td>
<td>131330</td>
<td>8376</td>
<td>55714</td>
<td>6195</td>
</tr>
<tr>
<td>SMT1</td>
<td>130568</td>
<td>9771</td>
<td>53935</td>
<td>6198</td>
</tr>
<tr>
<td>SMT2</td>
<td>127892</td>
<td>7943</td>
<td>51599</td>
<td>6131</td>
</tr>
</tbody>
</table>

Table 2: Annotations in VARTRA-SMALL
chunk type (lines 2b and 6b) information, as well as chunk (lines 2a, 2c, 6a and 6c) and sentence (lines 1 and 10) borders.

Table 3: Example queries to extract processual finite passive constructions

CQP also allows us to sort the extracted information according to the metadata: text registers and IDs or translation methods and tools. Table 4 shows an example of frequency distribution according to the metadata information. In this way, we can obtain data for our analyses of translation variation.

5 Preliminary Analyses

5.1 Profile of VARTRA-SMALL in terms of shallow features

We start our analyses with the comparison of translation variants only saved in our subcorpora: PT, CAT, RBMT, SMT1 and SMT2. The structure of the corpus, as well as the annotations available already allow us to compare subcorpora (translation variants) in terms of shallow features, such as type-token-ration (TTR), lexical density (LD) and part-of-speech (POS) distributions. These features are among the most frequently used variables which characterise linguistic variation in corpora, cf. (Biber et al., 1999) among others. They also deliver the best scores in the identification of translationese features. We calculate TTR as the percentage of different lexical word forms (types) per subcorpus. LD is calculated as percentage of content words and the percentages given in the POS distribution are the percentages of given word classes per subcorpus, all normalised per cent. The numerical results for TTR and LD are given in Table 5.
For the analysis of POS distribution, we decide to restrict them to nominal and verbal word classes. Tables 6 and 7 illustrate distribution of nominal – nouns, pronouns (pron), adjectives (adj) and adpositions (adp), and verbal word classes – verbs, adverbs (adv) and conjunctions (conj) – across different translation variants.

Table 6: Nominal word classes in % in VARTRA-SMALL

<table>
<thead>
<tr>
<th>subc</th>
<th>noun</th>
<th>pron</th>
<th>adj</th>
<th>adp</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>27.18</td>
<td>8.23</td>
<td>9.38</td>
<td>8.31</td>
<td>53.10</td>
</tr>
<tr>
<td>CAT</td>
<td>24.80</td>
<td>8.53</td>
<td>8.08</td>
<td>9.52</td>
<td>50.93</td>
</tr>
<tr>
<td>RBMT</td>
<td>24.80</td>
<td>8.61</td>
<td>8.91</td>
<td>9.01</td>
<td>51.32</td>
</tr>
<tr>
<td>SMT1</td>
<td>27.18</td>
<td>8.04</td>
<td>8.67</td>
<td>9.02</td>
<td>52.89</td>
</tr>
<tr>
<td>SMT2</td>
<td>29.78</td>
<td>7.28</td>
<td>10.42</td>
<td>8.64</td>
<td>56.11</td>
</tr>
</tbody>
</table>

Table 7: Verbal word classes in % in VARTRA-SMALL

<table>
<thead>
<tr>
<th>subc</th>
<th>verb</th>
<th>adv</th>
<th>conj</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>11.80</td>
<td>3.95</td>
<td>5.32</td>
<td>21.06</td>
</tr>
<tr>
<td>CAT</td>
<td>13.58</td>
<td>3.69</td>
<td>5.83</td>
<td>23.10</td>
</tr>
<tr>
<td>RBMT</td>
<td>12.90</td>
<td>2.74</td>
<td>6.34</td>
<td>21.99</td>
</tr>
<tr>
<td>SMT1</td>
<td>11.88</td>
<td>2.81</td>
<td>6.32</td>
<td>21.02</td>
</tr>
<tr>
<td>SMT2</td>
<td>9.09</td>
<td>2.52</td>
<td>6.06</td>
<td>17.67</td>
</tr>
</tbody>
</table>

Figure 1: Example of an annotated sample from VARTRA-SMALL

5.2 Interpretation of results

According to Biber (1999), high proportion of variable lexical words in a text is an indicator of richness and density of experiential meanings. This characterises the field of discourse (see section 2.2 above), and TTR, thus, indicates informational density. In terms of translationese (see section 2.1), TTR reveals simplification features of translations. Translations always reveal lower TTR and LD than their originals, cf. (Hansen, 2003).

The highest TTR, thus, the most lexically rich translation variant in VARTRA is the one produced by human translators: PT > RBMT > SMT2 > SMT1 > CAT. It is interesting that the other human-produced variant demonstrates the lowest lexical richness which might be explained by the level of experience of translators (student
Another reason could be the strength of pronominal cohesion and less explicit specification of domains. However, the comparison of the distribution of pronouns (devices for pronominal cohesion) does not reveal big differences between PT and CAT, cf. Table 6.

Another simplification feature is LD, which is also the lowest in CAT-subcorpus of VARTRA: PT > SMT2 > SMT1 > RBMT > CAT. Steiner (2012) claims that lower lexical density can indicate increased logical explicitness (increased use of conjunctions and adpositions) in translations. CAT does demonstrate the highest number of adpositions in the corpus, although the difference across subcorpora is not high, see Table 6.

The overall variation between the subcorpora in terms of TTR and LD is not high, which can be interpreted as indicator of levelling out (see section 2.1 above): translations are often more alike in terms of these features than the individual texts in a comparable corpus of source or target language.

In terms of nominal vs. verbal word classes, there seems to be a degree of dominance of nominal classes (56.11% vs. 17.67%) in SMT2 resulting in a ratio of 3.18 compared to other subcorpora, cf. Table 8.

<table>
<thead>
<tr>
<th>subc</th>
<th>nominal vs. verbal</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>53.10 : 21.06</td>
<td>2.52</td>
</tr>
<tr>
<td>CAT</td>
<td>50.93 : 23.10</td>
<td>2.20</td>
</tr>
<tr>
<td>RBMT</td>
<td>51.32 : 21.99</td>
<td>2.33</td>
</tr>
<tr>
<td>SMT1</td>
<td>52.89 : 21.02</td>
<td>2.52</td>
</tr>
<tr>
<td>SMT2</td>
<td>56.11 : 17.67</td>
<td>3.18</td>
</tr>
</tbody>
</table>

Table 8: Proportionality of nominal vs. verbal opposition in VARTRA-SMALL

The greatest contributors to this dominance are nouns and adjectives (Table 6 above). For CAT, we again observe the lowest numbers (the lowest noun vs. verb ratio) which means that this translation variant seems to be the most “verbal” one. According to Steiner (2012), German translations are usually more verbal than German originals. Comparing German and English in general, the author claims that German is less “verbal” than English. Thus, a higher “verbality” serves as an indicator of “shining though” (see 2.1 above), which we observe in case of CAT. However, to find this out, we would need to compare our subcorpora with their originals, as well as the comparable German originals.

5.3 First statistical experiments

We use the extracted shallow features for the first steps in feature evaluation. As our aim is to investigate the relations between the observed feature frequencies and the respective translation variants, we decide for correspondence analysis, a multivariate technique, which works on observed frequencies and provides a map of the data usually plotted in a two dimensional graph, cf. (Baayen, 2008).

As input we use the features described in 5.1 above: TTR, LD, nouns, adjectives (adj), adpositions (adp), verbs, adverbs (adv), conjunctions (conj). Additionally, we divide the class of pronouns into two groups: personal (pers.P) and demonstrative (dem.P) – devices to express pronominal cohesion. We also extract frequency information on modal verbs which express modality.

The output of the correspondence analysis is plotted into a two dimensional graph with arrows representing the observed feature frequencies and points representing the translation variants. The length of the arrows indicates how pronounced a particular feature is. The position of the points in relation to the arrows indicates the relative importance of a feature for a translation variant. The arrows pointing in the direction of an axis indicate a high contribution to the respective dimension. Figure 2 shows the graph for our data.

In Table 9, we present the Eigenvalues calculated for each dimension to assess how well our data is represented in the graph. We are able to obtain a relatively high cumulative value by the first two dimensions (representing x and y-axis in Figure 2), as they are the ones used to plot the two-dimensional graph. The cumulative value for the first two dimensions is 94.3%, which indicates that our data is well represented in the graph.

If we consider the y-axis in Figure 2, we see that there is a separation between human and machine translation, although SMT2 is on the borderline. CAT is also closer to MT, as it is plotted much closer to 0 than PT. Conjunctions, personal pronouns and adverbs seem to be most prominent contributors to this separation, as their arrows are...
the longest ones, and they point in the direction of the y-axis.

Verbs, adjectives and nouns seem to be most prominent contributors to the other division (considering the x-axis). Here, we can observe three groups of subcorpora: CAT and RBMT share certain properties which differ them from SMT2. PT remains on the borderline, whereas SMT1 tend slightly to SMT2.

### 6 Conclusion and Future Work

In this paper, we presented a comparable corpus of translations from English into German, which contains multiple variants of translation of the same texts. This corpus is an important resource for the investigation of variation phenomena reflected in linguistic features of translations. The corpus architecture allows us to extract these features automatically. Our preliminary results show that there are both similarities and differences between translation variants produced by humans and machine systems. We expect even more variation, if we compare the distribution of these features across text registers available in all subcorpora.

However, there is a need to inspect the reasons for this variation, as they can be effected by translator experience, restrictions of the CAT system applied or the training material used in MT.

We believe that our resources, as well as our research results will find application not only in contrastive linguistics or translation studies. On the one hand, our corpus provides a useful dataset to investigate translation phenomena and processes,
but on the other, it can be used for the development, optimisation and evaluation of MT systems, as well as CAT tools (e.g. translation memories).

In the future, we aim at expanding it with more data: (1) more texts for the existing registers (each register should contain around 30,000 words), (2) further text registers (e.g. academic, web and news texts). We also plan to produce further human and machine-generated translations, i.e. (3) machine translations post-edited by humans, as well as translation outputs of (4) further MT systems. Moreover, we aim at adding translations from German into English to trace variation influenced by language typology.

As the automatic tagging of part-of-speech and chunk information might be erroneous, we plan to evaluate the output of the TreeTagger and compare it with the output of further tools available, e.g. MATE dependency parser, cf. (Bohnet, 2010). Furthermore, the originals will be aligned with their translations on word and sentence level. This annotation type is particularly important for the analysis of variation in translation of certain lexico-grammatical structures.

A part of the corpus (CAT, RBMT and SMT subcorpora) will be available to a wider academic public, e.g. via the CLARIN-D repository.

Acknowledgments

The project “VARTRA: Translation Variation” was supported by a grant from Forschungsausschuß of the Saarland University. We are especially grateful to Anne-Katrin Schumann, Elke Teich and Noam Ordan for their comments. Also, we wish to thank the anonymous reviewers for their suggestions for improving our paper. All remaining errors remain ours.

References


85


