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[1] 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 specifc (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.

[1] CAT = computer-aided translation

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 77–86,
Sofia, Bulgaria, August 8, 2013. ©2013 Association for Computational Linguistics

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 assesment.

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 activitiy 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 profesionals (PT) were exported from the already existing corpus CroCo mentioned in section 1 above. The CAT variant was produced by student assistents who used the CAT tool ACROSS in the translation process[2]. The current RBMT variant was translated with SYSTRAN (RBMT1)[3], although we plan to expand it with a LINGUATEC-generated version[4]. For SMT, we have compiled two versions – the one produced with Google Translate[5] (SMT1), and the other one with a Moses system (SMT2).

Each translation variant is saved as a subcorpus and covers seven registers of written language: political essays (ESSAY), fictional texts (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), letters of share-holders (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 (Schmid, 1994).

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

subc	token	lemma	chunk	sent
PT	132609	9137	55319	6525
CAT	139825	10448	58669	6852
RBMT	131330	8376	55714	6195
SMT1	130568	9771	53935	6198
SMT2	127892	7943	51599	6131

Table 2: Annotations 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

[2] `www.my-across.net`

[3] SYSTRAN 6

[4] `www.linguatec.net`

[5] `http://translate.google.com/`

	EO	**PT**	**CAT**	**RBMT**	**SMT1**	**SMT2**
ESSAY	15537	15574	15795	15032	15120	14746
FICTION	11249	11257	12566	11048	11028	10528
INSTR	20739	21009	19903	20793	20630	20304
POPSCI	19745	19799	22755	20894	20353	19890
SHARE	24467	24613	24764	22768	22792	22392
SPEECH	23308	23346	24321	23034	22877	22361
TOU	17564	17638	19721	17761	17768	17671
TOTAL	132609	133236	139825	131330	130568	127892

Table 1: Tokens per register 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.

	query block	**example**
1.	<s>	
2a.	<chunk>	
2b.	[_.chunk_type="NC"]+	*Ein Chatfenster*
2c.	</chunk>	
3a.	[pos="VAFIN"&	
3b.	lemma="werden"]	*wird*
4.	[word!="."]*	*daraufhin*
5.	[pos="V.*PP"];	*angezeigt*
6a.	<chunk>	
6b.	[_.chunk_type="NC"]+	*das Transportgut*
6c.	</chunk>	
7.	[word!="."]*	*nicht*
8.	[pos="V.*PP"]	*akzeptiert*
9a.	[pos="VAFIN"&	
9b.	lemma="werden"]	*wird*
10.	</s>	

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.

method	**tool**	**register**	**freq**
CAT	Across	POPSCI	101
CAT	Across	SHARE	90
CAT	Across	SPEECH	89
CAT	Across	INSTR	73
RBMT	SYSTRAN	SHARE	63
RBMT	SYSTRAN	POPSCI	62
CAT	Across	TOU	58

Table 4: Example output of V2 processual passive across translation method, tool and text register (absolute frequencies)

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.

subc	**TTR**	**LD**
PT	15.82	48.33
CAT	14.10	44.60
RBMT	15.04	45.08
SMT1	14.32	46.03
SMT2	14.68	47.86

Table 5: TTR and LD in VARTRA-SMALL

```
<translation method="CAT" tool="Across" sourceLanguage="English">
<text "CAT_ESSAY_001.txt" register="ESSAY">
<s>
<chunk type="NC">
Die                  ART        d
weltweiten           ADJA       weltweit
Herausforderungen    NN         Herausforderung
</chunk>
<chunk type="PC">
im                   APPRART    im
Bereich              NN         Bereich
</chunk>
<chunk type="NC">
der                  ART        d
Energiesicherheit    NN         Energiesicherheit
</chunk>
<chunk type="VC">
erfordern            VVFIN      erfordern
</chunk>
<chunk type="PC">
über                 APPR       über
einen                ART        ein
Zeitraum             NN         Zeitraum
</chunk>
<chunk type="PC">
von                  APPR       von
vielen               PIAT       viel
Jahrzehnten          ADJA       jahrzehnte
nachhaltige          ADJA       nachhaltig
Anstrengungen        NN         Anstrengung
</chunk>
<chunk type="PC">
auf                  APPR       auf
```

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

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.

subc	noun	pron	adj	adp	total
PT	27.18	8.23	9.38	8.31	53.10
CAT	24.80	8.53	8.08	9.52	50.93
RBMT	24.80	8.61	8.91	9.01	51.32
SMT1	27.18	8.04	8.67	9.02	52.89
SMT2	29.78	7.28	10.42	8.64	56.11

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

subc	verb	adv	conj	total
PT	11.80	3.95	5.32	21.06
CAT	13.58	3.69	5.83	23.10
RBMT	12.90	2.74	6.34	21.99
SMT1	11.88	2.81	6.32	21.02
SMT2	9.09	2.52	6.06	17.67

Table 7: Verbal word classes in % in 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 translators). 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.

subc	**nominal vs. verbal**	**ratio**
PT	53.10 : 21.06	2.52
CAT	50.93 : 23.10	2.20
RBMT	51.32 : 21.99	2.33
SMT1	52.89 : 21.02	2.52
SMT2	56.11 : 17.67	3.18

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[6]. 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

[6] 'dim' lists dimensions, 'value' – Eigenvalues converted to percentages of explained variation in '%' and calculated as cumulative explained variation with the addition of each dimension in 'cum'.

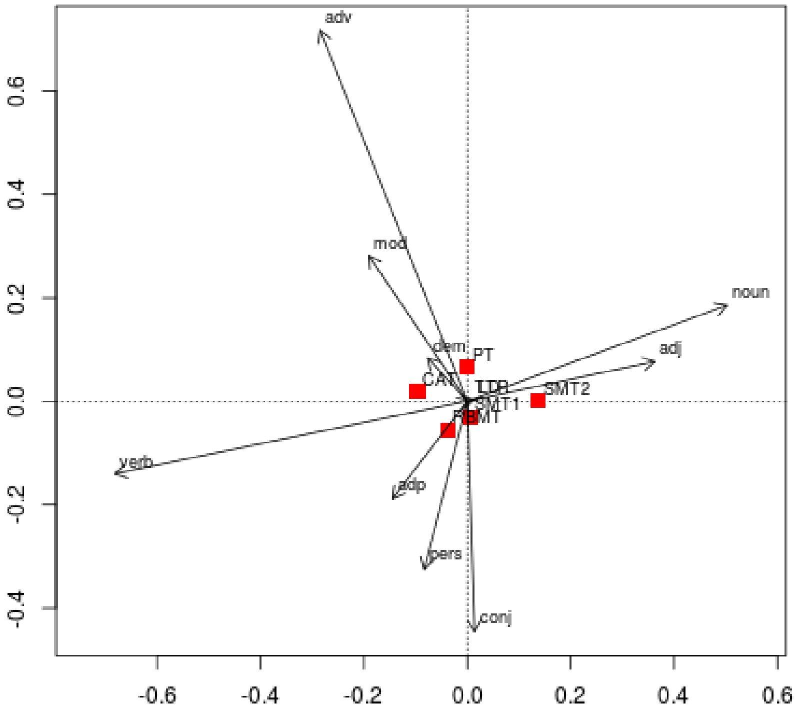

Figure 2: Graph for correspondence analysis on translation variants

dim	value	%	cum%	scree plot
1	0.005939	73.0	73.0	*************************
2	0.001726	21.2	94.3	*******
3	0.000352	4.3	98.6	*
4	0.000114	1.4	100.0	
	———	——		
Total:	0.008131	100.0		

Table 9: Contribution of dimensions

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

Across Personal Edition: Free CAT Tool for Freelance Translators. `http://www.my-across.net/en/translation-workbench.aspx`.

Harald Baayen. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Cambridge University Press.

Bogdan Babych and Anthony Hartley. 2004. Modelling legitimate translation variation for automatic evaluation of MT quality. *Proceedings of LREC-2004*, Vol. 3.

Mona Baker. 1995. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*, 7(2):223–43.

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. Harold Somers (ed.). Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager. Amsterdam and Philadelphia: Benjamins: 175–186.

Marco Baroni and Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21 (3): 259–274.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3): 209–226.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Douglas Biber. 1995. *Dimensions of Register Variation. A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Longman, London.

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. *The 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.

Susan Conrad and Douglas Biber (eds.). 2001. *Variation in English: Multi-Dimensional studies*. Longman, London.

The IMS Open Corpus Workbench. 2010. http://www.cwb.sourceforge.net

Stefan Evert. 2005. *The CQP Query Language Tutorial*. IMS Stuttgart, CWB version 2.2.b90.

Google Translate. Accessed July 2012. http://translate.google.com

Michael A.K. Halliday. 1985. *Spoken and written language*. Deakin University Press, Victoria.

Michael A.K. Halliday, and Riquaya Hasan. 1989. *Language, context and text: Aspects of language in a social semiotic perspective*. Oxford University Press.

Michael A.K. Halliday. 2004. *An Introduction to Functional Grammar*, 3. edition. Hodder Education.

Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner. 2013. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. Berlin, New York: de Gruyter.

Silvia Hansen. 2003. *The Nature of Translated Text – An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations*. Ph.D. Theses.

Juliane House. 1997. *Translation Quality Assessment: A Model Revisited*. Ph.D. Thesis.

Stig Johansson. Towards a multilingual corpus for contrastive analysis and translation studies. *Language and Computers*, 43 (1): 47–59.

Adam Kilgariff. 2010. Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project. *BUCC, 6th Workshop on Building and Using Comparable Corpora*, Valletta, Malta.

Phillip Koehn. 2005 Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit*.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL11)*.

David Kurokawa, Cyril Goutte and Pierre Isabelle. 2009. Automatic Detection of Translated Text and its Impact on Machine Translation. *Proceedings of MT-Summit-XII*.

Ekaterina Lapshinova-Koltunski, Kerstin Kunz and Marilisa Amoia. 2012. Compiling a Multilingual Spoken Corpus. *Proceedings of the VIIth GSCP International Conference : Speech and Corpora*. Firenze : Firenze University Press.

Sara Laviosa. 1997. How Comparable Can 'Comparable Corpora' Be? *Target*, 9(2): 289–319.

Gennady Lembersky, Noam Ordan and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*.

Linguatec Personal Translator 14. http://www.linguatec.net/products/tr/pt

Christian M.I.M. Matthiessen. 2001. The environment of translation. Erich Steiner and Colin Yallop (eds). *Exploring Translation and Multilingual Text Production: Beyond Content*. Berlin and New York: Mouten de Gruyter.

Christian M.I.M. Matthiessen. 2006. Frequency profiles of some basic grammatical systems: an interim report. Geoffrey Thompson and Susan Hunston (eds). *System and Corpus: Exploring connections*. Equinox, London.

Tony McEnery. 2003. *Oxford Handbook of Computational Linguistics*, chapter Corpus Linguistics: 448–463. Oxford: Oxford University Press.

Stella Neumann. 2011. *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin and New York: de Gruyter.

Kishore Papineni, Salim Roukus, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318.

Maja Popović and Aljoscha Burchardt. 2011. From Human to Automatic Error Classification for Machine Translation Output. *15th International Conference of the European Association for Machine Translation (EAMT 11)*.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *International Conference on New Methods in Language Processing*, Manchester (UK): 44–49.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. *Proceedings of the 15th Conference of the European Association for Machine Translation*: 73–80.

Lucia Specia, Najeh Hajlaoui, Catalina Hallett and Wilker Aziz. 2011. Predicting machine translation adequacy. *Machine Translation Summit XIII (2011)*: 19–23.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, Italy, 24-26 May 2006.

Erich Steiner. 2001. Translations English-German: Investigating the Relative Importance of Systemic Contrasts and of the Text Type translation. *SPRIKreports* 7:1–49.

Erich Steiner. 2004. *Translated texts: Properties, Variants, Evaluations*. Frankfurt a.Main: Peter Lang.

Erich Steiner. 2012. A characterization of the resource based on shallow statistics. Hansen-Schirra, Silvia, Stella Neumann and Erich Steiner (eds). *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. Berlin, New York: de Gruyter.

SYSTRAN Enterprise Server 6. Online Tools User Guide.

Elke Teich. 2003. *Cross-linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Berlin and New York: Mouton de Gruyter.

John S. White. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Further Approaches. *Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas*, 193–205.