# Chapter 5

# Variation in translation: Evidence from corpora

Ekaterina Lapshinova-Koltunski

The present paper describes a corpus-based approach to study variation in translation in terms of translation features. We compare texts, which differ in the source/target texts (English vs. German), production types (original vs. translation) and method of translation (human, computer-aided = CAT, machine) in terms of a theoretically-motivated set of features. In this study, we decide for the features which can be easily obtained on the basis of automatic corpus annotations, i.e. tokens, lemmas and part-of-speech tags. Our results show that there is variation in the mentioned translations in terms of the features under analysis.

## 1 Introduction: Aims and Motivation

In this paper, we apply corpus-based methods to analyse translation variants – translations from English into German produced with different translation methods.

Although numerous studies on translation operate with corpus-based methods, most of them concentrate on the questions concerning the nature of translations and their specific features, (e.g. Baker 1993; 1995; Laviosa 2002; Chesterman 2004) and others. The majority of them tried to generalise translation by defining certain rules or regularities of translated texts. Moreover, they mostly compare translations with originals, i.e. differences or similarities between translations and their source texts or comparable non-translated texts, ignoring variation which can be observed in different translation variants. Corpus-based studies dedicated to the analysis of variation phenomena involving translations, (e.g. Teich 2003; Steiner 2004; Neumann 2013), etc. concentrate on the analysis of human translations only. However, nowadays, translations are produced not only

Ekaterina Lapshinova-Koltunski. 2015. Variation in translation: evidence from corpora. In Claudio Fantinuoli & Federico Zanettin (eds.), *New directions in corpus-based translation studies*, 93–114. Berlin: Language Science Press

by humans but also with machine translation (MT) systems. Furthermore, new variants of translation appear due to the interaction of both, e.g. in computer-aided translation or post-editing.

In some works on machine translation the focus lies on comparing different translation variants, such as human vs. machine, as in (White 1994; Papineni et al. 2002; Babych & Hartley 2004; Popovic 2011). However, they all serve the task of automatic MT system evaluation and use the human-produced translations as references or training material only. None of them provide an analysis of specific linguistically motivated features of different text types translated with different translation methods, which is the aim of the present analysis.

In this study, we aim to apply corpus-based methods to prove the knowledge of translation features on a new dataset which contains different variants of translations, including human and machine translation.

The remainder of the paper is structured as follows. §2 presents studies we adopt as theoretical background for the selection of features under analysis. In §3.1, we describe the resources and methods used. In §4, we present the results of our analyses and their discussion, and in §5, we draw some conclusions and provide more ideas for future work.

## 2 Theoretical Background

Since the present study concentrates on the analysis of linguistic features of different translation variants, we address the existing studies on translation for their definition.

### 2.1 Related Feature Work

As already mentioned in §1 above, in most cases, these studies either analyse differences between original texts and translations (House 1997; Matthiessen 2001; Teich 2003; Hansen 2003; Steiner 2004), or concentrate on the properties of translated texts only (Baker 1995). Nevertheless, an important point is that most of them consider translations to have their own specific properties which distinguish them from the originals: both their source texts and comparable texts in the target language. These features establish the specific language of translations which is called *translationese* (Gellerstam 1986). Comparing Swedish translations from English with Swedish original texts, the author stated significant differences between them, whereas not all of them were attributable to the source language. This coincides with what Frawley (1984) called "third code", describ-

ing features of translational language which are supposed to be different from both source and target languages.

Later, Mona Baker emphasised general effects of the process of translation that are independent of source language, e.g. in Baker (1993; 1995). Analysing characteristic patterns of translations, she excluded the influence of the source language on a translation altogether. Within this context, she proposed *translation universals* – linguistic features which typically occur in translated rather than original texts. According to Baker (1993), they are independent of the influence of the specific language pairs involved in the process of translation. Other scholars (e.g. Toury 1995 or Chesterman 2004) operate with other terms – "laws" or "regularities". We prefer to use the term "translation features" or "phenomena" in the present study: to claim the features "universal" we would need to analyse more language pairs and translation directions, and to call them "laws" and "regularities", we would need to test more conditions, e.g. cognitive factors, status of translation, etc., which is not possible with the bilingual dataset at hand.

Translation features can be classified according to different parameters. For instance, Chesterman (2004) makes a distinction between *S-universals* and *T-universals*: the first comprises differences between translations and their source texts, and the second covers the differences between translations and comparable non-translated texts. A more fine-grained classification includes the following features: *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, *levelling out*– individual translated texts are more alike than individual original texts, in both source and target languages, and *interference* – features of the source texts are observed in translations. For the second last, we prefer the term convergence proposed by Laviosa (2002), which implies a relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features, e.g. lexical density, sentence length, etc. in contrast to originals. For the last feature, we also prefer to use the term *shining through* defined by Teich (2003).

All these features have been widely analysed in corpus-based translation studies for different language pairs, e.g. in Laviosa (1996) for English translations from a variety of source languages, in Mauranen (2000) for English–Finnish translations, in Teich (2003) for English and German translations, and others. Yet, all of them concentrate on human translations only.

Moreover, some recent corpus-based studies applied machine learning supervised methods to automatically differentiate between translations and originals

(e.g. Baroni & Bernardini 2006). These approaches found application in some recent works on natural language processing, e.g. those on cleaning parallel corpora obtained from the Web, or improvement of translation and language models in MT (e.g. Kurokawa, Goutte & Isabelle 2009; Koppel & Ordan 2011; Lembersky, Ordan & Wintner 2012).

We employ the knowledge from these studies, as well as techniques applied to explore the differences between translation variants under analysis, including the features related to their source texts as well as those of comparable target texts.

## 2.2 Translation Features and their Operationalisation

We group the features described above into three classes according to their correlations, especially in their operationalisation: 1) *simplification*, 2) *explicitation*, 3) *normalisation* vs. *shining through* and 4) *convergence*. Simplification can be analysed on different levels, e.g. lexical, syntactic or semantic. If core patterns of lexical use are observed (see Laviosa 1998), we can identify simplification comparing the proportion of content vs. grammatical words. Translated texts have a relatively low percentage of content words, and the most frequent words are repeated more often. This means, that both lexical density and type–token–ratio of translations are lower than those of their source texts and the comparable texts in the target language. Besides, more general terms are expected to be used in translations. On the level of syntax, one can observe short sentences which replace long ones and a lower average sentence length in general.

Explicitation involves the addition and specification of lexical and grammatical items, with the help of which implicit information in the source text is "spelled out" in its translation. The indicators of this feature include a higher ratio of function words which make grammatical relations explicit, specific terms replacing more general terms (the opposite of simplification), disambiguation of pronouns, increased use of cohesive devices, e.g. conjunctions, and others. In terms of cohesion, one would also expect more nominal (expressed with nominal phrases) than pronominal reference (expressed with personal pronouns) in translations.

Simplification and explicitation features correlate and may be just the opposite of each other. For example, if we observe more specific terms replacing general terms in translation, we face the feature of explicitation, and not simplification. Normalisation and "shining through" can also be measured on different levels, depending on the languages involved. Both features depend on the contrasts between these languages: normalisation implies the exaggerated use of the patterns typical for the target languages, whereas "shining through" involves the

patterns typical for the source language (but not specific for the target language) that can be observed in translations. For instance, normalisation can be verified by a great number of typical collocations and neutralised metaphoric expressions. Baker (1996) claims that influence of normalisation depends on the status of the source language: "the higher the status of the source text and language, the less the tendency to normalise". We assume that the languages with a higher status also tend to "shine through" more often. For example, if we analyse translations from English, we would probably observe more "shining through" than normalisation, as English has the highest world language status.

And finally, convergence is a homogeneity feature of translations: they reveal less variation if we compare them to original texts. Convergence can also be observed on all levels of a language system. In accordance with the convergence phenomenon, one would expect that the lexical, grammatical and syntactic features under analysis will reveal smaller differences in translations than in originals.

## 2.3 Hypotheses

For our analysis of translation variants, we select a set of operationalisation of the features described in §2.2 above.

1. **Simplification** - We expect that our translated texts have a lower percentage of content words vs. grammatical words than their English source texts and the comparable German texts. Also, words are repeated more often in translations. Thus, we observe lower lexical density and type–token–ratio in our translations. In the analysis of English to German translations, we exclude sentence length as operationalisation for simplification. Due to the systemic differences in the morphology, German sentences are generally shorter than those in English, as they contain one-word compounds. To measure this uniformly, we need to split compounds and measure their parts as tokens, which is not feasible within this study.

2. **Explicitation** - Our translated texts reveal more cohesive explicitness than English and German originals: we can observe more conjunctions, less pronominal reference and less general nouns in translations than in English and German originals.

3. **Normalisation/ shining through** - If the translations under analysis demonstrate features more typical for English than for German, we observe "shining through". If there are more features typical for German originals,

then our translations demonstrate normalisation. Here, we use the knowledge from contrastive analysis, e.g. German–English contrasts described in Hawkins (1986), König & Gast (2007), Steiner (2012). For example, we know that English is more "verbal" than German. This can be proved by comparing the distribution of nominal and verbal phrases in both translations and originals. English originals are expected to contain more verbal than nominal phrases. The phenomenon of "shining through" will be confirmed in our data if translations contain more verbal phrases than German originals. On the contrary, if they contain less verbal phrases than German originals, the normalisation hypothesis will be confirmed.

4. **Convergence** - The variation of the features in 1 to 3 is not great if we compare translation variants: they are similar to each other, i.e. the features are distributed homogeneously.

## 3 Resources, Methods and Tools

To prove the hypotheses formulated in §2.3, we need to compare the distribution of the features under analysis across translation variants, their English sources as well as comparable German originals. For this, we analyse frequency distribution information of lexico-grammatical patterns which serve as operationalisation for these features. The patterns are extracted from a corpus at hand, and evaluated with univariate statistical methods (e.g. significance analysis).

### 3.1 Corpus Resources

For our investigations, we use VARTRA-SMALL, (see Lapshinova-Koltunski 2013), a translation corpus which contains German translation variants from English produced with different translation methods: by (1) human professionals (PT), (2) human inexperienced translators (CAT), with (3) rule-based MT systems (RBMT) and (4) two statistical MT systems (SMT1 and SMT2). Translations by professionals (PT) were exported from the already existing corpus CroCo (Hansen-Schirra, Neumann & Steiner 2013). The same corpus provides source English texts (EO) and comparable German originals (GO). Thus, we can compare source English texts with their multiple translations into German, as well as to comparable German originals.

The CAT variant was produced by trained translators with at least BA degree, who have no/little experience in translation. All of them applied computer-aided

tools while translating the given texts.[1] The rule-based machine translation variant was translated with SYSTRAN (RBMT),[2] whereas for statistical machine translation we have two further versions – the one produced with Google Translate[3] (SMT1), and the other – with a self-trained Moses system (SMT2) (see Koehn et al. 2007).

The analysed dataset covers seven registers of written language: political essays (ESSAY), fictional texts (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), "letters to share-holders" (SHARE), prepared political speeches (SPEECH), and tourism leaflets (TOU). The size of all translation variants in VARTRA-SMALL comprises approx. 600 thousand tokens. The subcorpora of originals from CroCo comprise around 250 thousand words each.

All subcorpora under analysis are tokenised, lemmatised and tagged with part-of-speech information, segmented into syntactic chunks and sentences. The annotations of the VARTRA-SMALL subcorpora were obtained with Tree Tagger (see Schmid 1994). The availability of these annotation levels in both corpora allows us to analyse certain lexico-grammatical patterns – operationalisation of the translation features under analysis, defined in §2.3.

The subcorpora are encoded in CWB format (CWB, 2010) and can be queried with the help of the CQP regular expressions described in Evert (2005).

Alignment on sentence level is available for professional translations only: each translation is aligned with its English source on sentence level. No alignment is provided for further translation variants at the moment. However, this annotation level is not necessary for the extraction of the operationalization used in the present paper.

### 3.2 Feature Extraction

As already mentioned in §3.1 above, the corpus at hand can be queried with CQP, which allows the definition of language patterns in form of regular expressions based on string, part-of-speech and chunk tags as well as further constraints.

To prove the hypothesis for simplification indicated by lexical density (proportion of content words), we extract information on the distribution of content words in our corpus, for which the query 1 in Table 1 is used.

To extract the corpus evidence of explicitation, we apply queries 2 to 5. Query 2 is used to extract all occurrences of coordinating and subordinating conjunctions,

[1] We used the open source tool ACROSS, see www.my-across.net.

[2] SYSTRAN 6.

[3] http://translate.google.com/.

Table 1: Queries for feature extraction

| | query element | explanation |
|---|---|---|
| 1 | [pos="VV.*\|N.*\|ADJ.*\|ADV"] | a full verb/noun or an adjective/adverb |
| 2 | [pos="KON\|KOUS"] | connector or subordinator |
| 3 | <NP>[pos="PPE.*"]+</NP> | nominal phrase filled with a pronoun |
| 4 | <NP>[ ]+</NP> | any nominal phrase |
| 5 | [lemma=RE($general)] | nouns from a list |
| 6 | (<NP>[]+</NP>)\|(<PP>[]+</PP>) | nominal phrase or prepositional phrase |
| 7 | <VP>[]+</VP> | verbal phrase |

whereas queries 3 and 4 are used for extraction of information on pronominal vs. nominal reference in the corpus.

We calculate this as proportion of nominal phrases filled with personal pronouns (query 3) to all nominal phrases in the corpus (query 4). Query 5 is used to extract occurrences of general terms in order to compare their proportion to all nouns in the dataset. For this, we use a simple lexical search – we extract a closed class of lexical items of which we know the members. Here, we use lists of general nouns as defined in (Dipper, Seiss & Zinsmeister 2012). For normalisation/shining through, we extract all occurrences of nominal and prepositional phrases (query 6) vs. verbal phrases (query 7). Convergence is proved with the help of all patterns described above.

As we operate with low-level features which do not require formulation of complex lexico-grammatical patterns, we believe that our feature extraction procedures are adequate for the present task. Its only shortcoming is the potential noise caused by tagging errors, especially in case of machine translation. In the latter, we observe a number of untranslated words which are tagged as named entities by automatic part-of-speech taggers. In the longer run, we aim to include deeper structures into our analysis which would require parsed data.

## 4 Results and their Interpretation

### 4.1 Simplification

In the first step, we want to test if lexical density and type–token–ratio are lower in translation variant than in EO and GO.

Table 2: STTR and LD in VARTRA-SMALL

| | EO | GO | PT | CAT | HU-$\bar{x}$ | RBMT | SMT1 | SMT2 | MT-$\bar{x}$ | Trans-$\bar{x}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| LD | **45.72** | **45.49** | 46.23 | 44.60 | **45.64** | 45.08 | 46.02 | 47.86 | **46.30** | **45.97** |
| STTR | **367.5** | **369.9** | 360.8 | 336.4 | **348.6** | 335.2 | 350.4 | 309.0 | **331.5** | **338.4** |

As already mentioned above, lexical density (LD in Table 2) is measured as a proportion of all content words in our corpus. Unexpectedly, average lexical density in translations (Trans-$\bar{x}$ in Table 2) does not differ from that of both source and comparable originals. Moreover, if we consider the mean values for human and machine translations separately (HU-$\bar{x}$ and MT-$\bar{x}$ respectively); the latter demonstrates even higher LD than human translations and English and German originals. The lowest figure is obtained for CAT, which demonstrates a value below the average. The highest value is observed for SMT2 (47.86).

We explain this by the lexical constraints of the Moses-based system: this system depends on the parallel data used for its training. If the parallel data does not contain translations for some words in a text to be translated, the system keeps them untranslated. In the automatic part-of-speech tagging, these words are then tagged as proper nouns (NE) which leads to their high amount in texts, as seen in example (1).

However, the overall difference between originals and translations is not great, which means that lexical density is not an indicator of simplification in our dataset, as the translated texts show an amount of content words similar to that of the source and comparable originals.

(1) Wenn Sie strongly , believe , wie ich , dass Großbritannien
KOUS PPER NE $ NE $ KOKOM PPER $ KOUS NE
einen zentralen Platz einnehmen müssen in Europe's
ART ADJA NN VVINF VMINF APPR NE
decision-making…
NE

Another indicator of simplification is type–token–ratio which we measure as standardised type–token–ratio (STTR) – a percentage of different lexical word forms (types) per text. As expected, on average, translations show lower STTR than their source texts and comparable originals, see Table 2. Mean value of human translations is also higher than that of machine (348.6 vs. 338.6 respectively). Within translations, the highest STTR, thus, the most lexically rich translation

variant in our corpus, is the one produced by professional human translators (360.8), followed by SMT1 (350.4), and CAT (336.4). The level of the latter is close to the average of all translations but lower than that of human translations. The lowest figure is obtained for SMT2 (309.0). This can once again be explained by the fact that this translation variant contains a great deal of untranslated English words, the lemmas of which cannot be identified by the lemmatiser and thus is replaced with "<unknown>", see example (2).

(2) Closing die Gap Zwischen *Supply* und Die Nachfrage
<*unknown*> d <*unknown*> zwischen <*unknown*> und d Nachfrage
nach A *balanced* , umfassende Energiepolitik ist dringend
nach A <*unknown*> , umfassend Energiepolitik sein dringend
erforderlich , die langfristige Stärke der amerikanische wirtschaftlichen
erforderlich , d langfristig Stärke d amerikanisch wirtschaftlich
und nationalen.
und national <*unknown*>

Interestingly, student translations are closer to the RBMT translation variant in terms of both STTR (336.4 vs. 335.2) and LD (44.60 vs. 45.08). Analysing human and machine translation separately, we observe the same ranking in terms of both indicators: PT > CAT, whereas it is not stable in machine translation: while SMT2 ranks first in LD, it occupies the last position with its STTR value.

## 4.2 Explicitation

To analyse this feature in our corpus, we measure cohesive explicitness in all subcorpora. Here, we calculate the relative frequencies for conjunctions (conj in Table 3, normalised to the total number of words per thousand), proportion of nominal phrases filled with pro-forms vs. full nominal phrases (pronNP in Table 3, normalised per thousand), as well as proportion of general nouns vs. all noun occurrences (gennoun in Table 3 normalised per thousand) in translations and English and German originals.

According to our hypothesis in §2.3, we expect more conjunctions, less pronominal reference and less general nouns in translations than in originals. If we compare the values of all translations (Trans-$\bar{x}$) with those of their originals, our hypothesis can be confirmed for pronominal reference and general nouns only: Trans-$\bar{x}$ (137.76) < EO (204.67) and Trans-$\bar{x}$ (20.51) < EO (48.71). Translations demonstrate a lower and not higher distribution of conjunctions, Trans-$\bar{x}$ (50.67) < EO (53.80), contrary to what was expected. If we consider human and machine

Table 3: Explicitation indicators

| | conj | pronNP | gennoun |
|---|---|---|---|
| EO | **53.80** | **204.67** | **48.71** |
| GO | **43.58** | **127.14** | **23.85** |
| PT | 47.58 | 232.76 | 19.64 |
| CAT | 49.67 | 139.12 | 19.93 |
| HU-$\bar{x}$ | **48.33** | **184.67** | **19.74** |
| RBMT | 53.32 | 144.46 | 23.18 |
| SMT1 | 52.54 | 143.15 | 21.22 |
| SMT2 | 53.69 | 39.85 | 19.46 |
| MT-$\bar{x}$ | **53.18** | **107.65** | **21.19** |
| Trans-$\bar{x}$ | **50.76** | **137.76** | **20.51** |

translation separately, we see that values for machine translation are much closer to EO (53.18 vs. 53.80), which means that in these translation variants, cohesive relation expressed via conjunctions, were preserved similarly to their English originals. Conversely, fewer conjunctions were used in human translation. The number is still higher than observed in German originals (43.58); therefore, we cannot assume the phenomenon of normalization here. This means that, in our dataset, human translators tend to keep that relation implicit, as seen in example (3).

(3) a. Negative molecules moved into the nurse cells if the egg was made negative, while positive molecules stayed put (EO-POPSCI).

b. Wenn das Ei auch negativ war, bewegten sich negativ geladene Moleküle in die Nährzellen, positiv geladene Moleküle blieben an Ort und Stelle (PT-POPSCI).

Admittedly, our extractions exclude occurrences of adverbial conjunctions (as we extract coordinating and subordinating conjunctions only). Previous analyses (e.g. Kunz & Lapshinova-Koltunski 2014) show that this syntactic type of conjunction is highly frequent in German. We suppose that English coordinating and subordinating conjunctions are in some cases translated with adverbials in German.

Example (4) extracted from our corpus demonstrates variants of translation of the English subordinating conjunction “while”. In both human translations (b.

and c.), conjunctive relation is transferred with an adverbial phrase. In machine translated variants (d. to f.), "while" is translated directly with *während*, so the type of cohesive conjunction is preserved as it was in the original.

(4) a. And *while* this will vary from quarter to quarter based on large cash outlays such as tax payments and end-of-year compensation payments, we were pleased with our average positive cash flow for the year from operations of $ 1.5 billion per quarter. (EO-SHARE).

b. Dieser Wert schwankt bei Betrachtung verschiedener Quartale. Dafür sind Auszahlungen hoher Beträge (z.B. Steuerzahlungen) sowie Ausgleichszahlungen am Jahresende verantwortlich. Mit dem durchschnittlichen operativen Cashflow von 1,5 Milliarden US-Dollar pro Quartal sind wir *jedoch* höchst zufrieden (PT-SHARE).

c. Dieser Cash Flow fällt zwar aufgrund von hohen Barauslagen, wie Steuern und Ausgleichszahlungen am Jahresende, in jedem Quartal unterschiedlich aus, *dennoch* waren wir mit unserem durchschnittlichen jährlichen Cash Flow aus laufenden Geschäftstätigkeiten von 1,5 Millionen $ pro Quartal zufrieden (CAT-SHARE).

d. Und basiert *während* dieses von Viertel zu das Viertel schwankt, das auf grossen Barauslagen wie Steuerzahlungen und Jahresendeausgleichszahlungen, wurden wir mit unserem durchschnittlichen positiven Cashflow für das Jahr von den Operationen von $1,5 Milliarde pro Viertel gefallen (RBMT-SHARE).

e. Und *während* dies von Quartal zu Quartal basierend auf grosse Barauslagen wie Steuer-Zahlungen und End-of-Jahres-Ausgleichszahlungen variieren, wurden wir mit unserer durchschnittlichen positiven Cashflow für das Jahr aus dem operativen Geschäft von 1,5 Milliarden Dollar pro Quartal (SMT1-SHARE).

f. Und *während* diese je nach Viertel bis Viertel auf der Grundlage grosse Geld ausgegeben wie Steuerzahlungen und abschliessende Entschädigung payments, freuen wir uns mit unseren durchschnittliche positive Cashflow für das Jahr von Maßnahmen der $1.5 Milliarden pro quarter (SMT2-SHARE).

In some cases, cohesion might be expressed with different cohesive devices in the two languages under analysis. For instance, the conjunction "while" in the

source sentence in example (5) is substituted with a reference expressed with the pronominal adverb *dabei* in PT, see (5-b). Pronominal adverbs expressing a reference are typical for German and are rare in English. At the same time, we observe the adoption of the cohesive device used in the source sentence also in other translation variants (c. to f.).

(5) a. My father preferred to stay in a bathrobe and be waited on for a change *while* he lead the stacks of newspapers me and my grandmother saved for him (EO-FICTION).

b. Mein Vater ist lieber im Bademantel geblieben und hat sich zur Abwechslung mal bedienen lassen und *dabei* die Zeitungsstapel durchgelesen, die ich und meine Großmutter für ihn aufgehoben haben (PT-FICTION).

c. Mein Vater saß die ganze Zeit im Bademantel da und ließ sich zur Abwechslung bedienen, *während* er die Zeitungen laß, die meine Großmutter und ich für ihn aufgehoben hatten (CAT-FICTION).

d. Mein Vater bevorzugt, um in einem Bademantel zu bleiben und auf eine Änderung, *während* er die Stapel von Zeitungen ich und meine führt Großmutter an gewartet zu werden gerettet für ihn (RBMT-FICTION).

e. Mein Vater lieber im Bademantel bleiben und werden wartete auf eine Veränderung, *während* er die Stapel von Zeitungen mich und meine Großmutter für ihn gerettet führen (SMT1-FICTION).

f. My Vater lieber Aufenthalt in einem bathrobe und gewartet werden über einen Klimawandel, *während* er die Stapeln kramen Zeitungen für mich und meine Großmutter him (SMT2-FICTION).

In terms of reference, translations demonstrate less noun phrases filled with pronouns than their source texts in English: 137.76 (Trans-$\bar{x}$) vs. 204.14 (EO), whereas the opposite phenomenon is observed, if we compare them to the original texts in German. In this case, we observe more pronominal reference in translations than in comparable originals (137.76 vs. 127.14). However, variation is observed across translation variants: while in human translations pronominal reference is much higher and tends to the values of EO, machine translation shows values which are lower when compared to both EO and GO. This low value is obviously caused by the small amount of pronominal phrases in SMT2. Here, we suppose that many pronouns remained untranslated in certain registers, as seen in example (5-f) above, and were wrongly tagged in the part-of-speech an-

notation. Moreover, we observe a high number of pronominal references in PT (232.76), which contradicts the hypothesis in §2.3.

The figures obtained for general nouns confirm our hypothesis about their low frequency in translations. On average, translations demonstrate a lower amount of general nouns than EO and GO (20.51 vs. 48.71 and 23.85 respectively). RBMT is the only translation variant whose distribution of general nouns is similar to that of GO. As seen from the values for the originals, there are more general nouns in EO than in GO. This means that this type of nouns is more typical for English than for German. Hence, we observe normalisation in terms of general nouns in all translation variants of our corpus.

In the analysis of explicitation in translations from English into German, one should also take into account the fact that German is more explicit than English, which could also have influenced on the results obtained.

#### 4.2.1 Normalisation and "shining through"

To analyse normalisation and "shining through", we extracted all occurrences of nominal and prepositional phrases and compared them with the occurrences of verbal phrases. Table 4 demonstrates the proportions of nominal (nominal and prepositional phrases) and verbal (verbal phrases) classes across all subcorpora under analysis. As already mentioned in §2.3 above, German is less "verbal" than English, which is confirmed in our data: GO contains less verbal phrases than EO.

The mean value of verbal phrases for all translations comprises 28.63, which is much lower than that of GO. This indicates the phenomenon of normalisation in this case. Comparing the values across translation variants, we observe variation in the degree of normalisation – it is less pronounced in human than in machine translation (33.59 vs. 24.64 respectively). Moreover, human translations produced by professionals are very close to German originals in terms of the distribution of nominal vs. verbal phrases, which means that they demonstrate neither normalisation nor "shining through" if we consider the indicators under analysis.

The higher noun-verb-ratio (NVratio in Table 4) is observed for SMT2. The reason for it could once again be the erroneous part-of-speech tagging which results from the gaps in training data used for SMT2. Most untranslated verbs (e.g. promote, report, import) or verbal forms (recognising, closing, helping, etc.) were tagged as nouns or adjectives.

Overall, the results are rather surprising. Analysing examples in our corpus, we notice that source verbal phrases in human translations from English into German are often translated as nominal phrases, see examples (6-a), (6-b) and

Table 4: Proportionality of nominal vs. verbal opposition in VARTRA-SMALL

| subc | nominal | verbal | NVRATIO |
|---|---|---|---|
| EO | 59.45 | 40.55 | 1.47 |
| GO | 61.95 | 38.05 | 1.63 |
| PT | 61.92 | 38.08 | 1.63 |
| CAT | 71.87 | 28.13 | 2.56 |
| HU-$\bar{x}$ | 66.41 | 33.59 | 1.98 |
| RBMT | 72.42 | 27.58 | 2.63 |
| SMT1 | 74.38 | 25.62 | 2.90 |
| SMT2 | 79.54 | 20.46 | 3.89 |
| MT-$\boldsymbol{x}$ | 75.35 | 24.64 | 3.06 |
| **Trans-$\bar{x}$** | 71.36 | 28.63 | 2.49 |

(6-c). However, they are often left as verbal phrases in machine translation, as in examples (6-d), (6-e) and (6-f). Therefore, we would expect machine-produced translations to have a lower noun-verb-ratio, which is not the case in the quantitative data. To analyse the correspondences between source and target phrases we need to align our subcorpora, which is not available at the moment.

(6) a. Settings *changed* here override settings *changed* anywhere else (EO-INSTR).

b. Die hier vorgenommenen *Änderungen* setzen alle anderen *Änderungen* außer Kraft (PTINSTR).

c. Hier vorgenommene Einstellungs*änderungen* sind allen anderen Einstellungs*änderungen* übergeordnet (CAT-INSTR).

d. Die Einstellungen, die hier *geändert* werden, heben die Einstellungen auf, die irgendwoanders *geändert* werden (RBMT-INSTR).

e. Hier *geänderten* Einstellungen überschreiben Einstellungen, die anderswo *geändert* (SMT1-INSTR).

f. … bei dem Sie überhaupt hier über Rahmenbedingungen *geändert* Settings überall else (SMT2-INSTR).

### 4.3 Convergence

In our last hypothesis, we test if the analysed translations exhibit convergence – the variation of the features across translation variants in our corpus is not high. For this purpose, we consider the indicators analysed in §1, §2 and §3 above: STTR, LD, conj, pronNP, gennoun and NVratio. The overall variation between the subcorpora is relatively low for all features, except for pronominal reference and noun-verb-ratio (see Figure 1), which means that translation variants in our corpus are alike in terms of the features considered. Most prominent indicators for convergence are that of simplification. We remove pronominal reference and noun-verb-ratio from the data matrix and calculate p–values using the Pearson's chi–square test, which is a univariate statistical method to reveal significant differences between variables. If p–value is < 0.05, then the difference between the compared subcorpora (translation variants) is not significant.

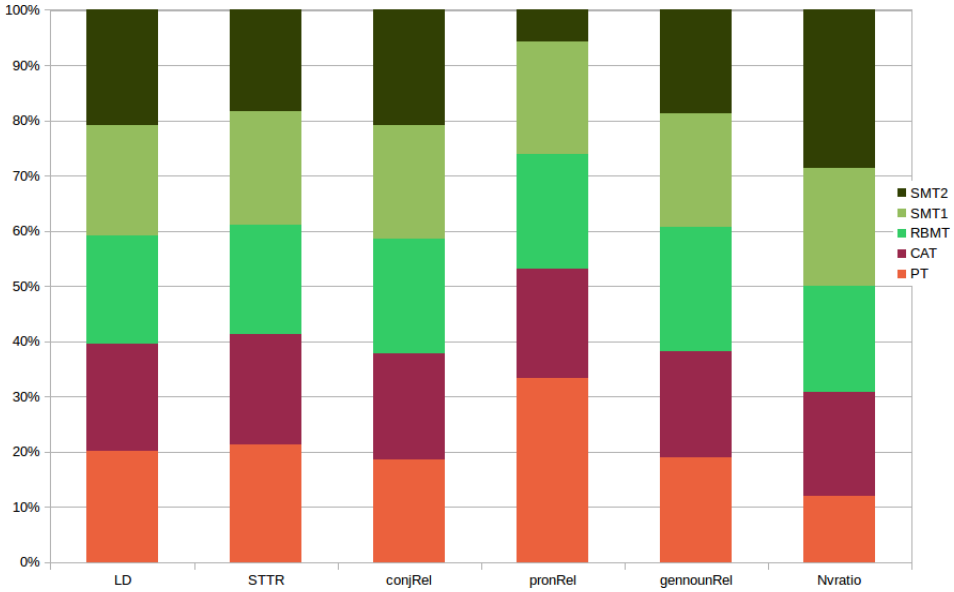


Figure 1: Levelling out in VARTRA-SMALL

We calculate p–value for all pairs of subcorpora in **VARTRA**. The results confirm our assumptions, as p–value is above 0.05 in almost all cases (see Table 5). An exception is pair **PT-SMT2**, where we observe a p–value of approx. 0.01. Our translation variants therefore converge, as expected, as there is no significant difference between almost all subcorpora; they are alike in terms of the analysed

Table 5: p–values for comparison of translation variants

| | subc | p–value |
|---|---|---|
| **PT** | vs. CAT | 0.8863 |
| | vs. RBMT | 0.5663 |
| | vs. SMT1 | 0.8806 |
| | vs. SMT2 | 0.0142 |
| **CAT** | vs. RBMT | 0.9307 |
| | vs. SMT1 | 0.9986 |
| | vs. SMT2 | 0.0980 |
| **RBMT** | vs. SMT1 | 0.9373 |
| | vs. SMT2 | 0.1771 |
| **SMT1** | vs. SMT2 | 0.0731 |

phenomena, which are indicators of simplification, explicitation and normalisation.

### 4.4 Summary

Summarising the obtained results, we found that not all hypotheses formulated in §2.3 above can be applied to our dataset. Both type–token–ratio as well as lexical density do not serve as good indicators of simplification in this case. In terms of explicitation, we should also think of further operationalisation, as those chosen reveal rather other phenomena (e.g. normalisation). The hypotheses about normalisation and "shining through" can be confirmed only in part and reflect high variations across translation varieties. The only assumption confirmed by our data is that of convergence. The analysed translation variants converge, as there is no significant difference between them in terms of the analysed phenomena.

## 5 Conclusion and future work

In this paper, we analysed translation variants produced by humans and machine systems and compared them to their English source texts, as well as comparable German originals. With the help of lexicogrammatical patterns, we were able

to trace differences and similarities between them, which indicate the following translation features: simplification, explicitation, normalisation and convergence. Although our analysis includes translations from English into German, we could not detect "shining through" – at least with the indicators at hand. The analysed features vary if we consider translation variants or their groups separately, e.g. in terms of explicitation or normalisation. At the same time, we observe convergence in translation, especially if we take simplification into account.

We believe that we should include more factors into the analysis to explain the variation observed. For example, in some cases, we should revise our hypotheses and their operationalisation, as contrasts between languages should be taken into account. We also need to look at the "experience" factor – this could verify the differences between two human translations observed for some features. Furthermore, restrictions of the translation memory applied in CAT or the training material used in SMT can also have an influence on the distribution of lexico-grammatical patterns. For this, a closer inspection of correlations between translation memory as well as applied SMT training material (parallel corpora) is required, which is planned for our future work.

We also plan to align originals with their translations on word and sentence level to allow analysis of certain phenomena involved, e.g. translation of ambiguous cases, direct translation solutions, see 4.3 and their multiple variants.

## Acknowledgments

The project *VARTRA: Translation Variation* was supported by a grant from Forschungsausschuss of the Saarland University.

## References

Babych, Bogdan & Tony Hartley. 2004. Extending the BLEU MT Evaluation Method with Frequency Weightings. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, 621–628.

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 233–250. Amsterdam: John Benjamins.

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7(2). 223–243.

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

Baroni, Marco & Silvia Bernardini. 2006. 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.

Chesterman, Andrew. 2004. Beyond the particular. In Anna Mauranen & Pekka Kujamäki (eds.), *Translation universals: Do they exist?*, 33–49. Amsterdam: John Benjamins.

Dipper, Stefanie, Melanie Seiss & Heike Zinsmeister. 2012. The use of parallel and comparable data for analysis of abstract anaphora in German and English. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eight international conference on language resources and evaluation (lrec 2012)*, 138–145. Paris: ELRA.

Evert, Stefan. 2005. *The CQP Query Language Tutorial*. CWB version 2.2.b90. http://www.ims.uni-stuttgart.de/forschung/projekte/CorpusWorkbench/CQPTutorial/cqp-tutorial.2up.pdf.

Frawley, William. 1984. Prolegomenon to a theory of translation. In William Frawley (ed.), *Translation: Literary, Linguistic and Philosophical Perspectives*, 159–175. London: Associated University Press.

Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. In Lars Wollin & Hans Lindquist (eds.), *Translation Studies in Scandinavia*, 88–95. Lund: CWK Gleerup.

Hansen, Silvia. 2003. *The nature of translated text: An interdisciplinary methodology for the investigation of the specific properties of translations* (Saarbrücken dissertations in computational linguistics and language technology). German Research Center for Artificial Intelligence, Saarland University.

Hansen-Schirra, Silvia, Stella Neumann & Erich Steiner. 2013. *Cross-linguistic corpora for the study of translations. Insights from the language pair English-German*. Berlin: de Gruyter.

Hawkins, John A. 1986. *A comparative typology of English and German*. London: Croom Helm.

House, Juliane. 1997. *Translation quality assessment. A model revisited*. Tübingen: Narr.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin & Evan Herbst. 2007.

Moses: Open source toolkit for statistical machine translation. In *Proceedings of acl-2007*, 177–180. Prague.

Koppel, Moshe & Noam Ordan. 2011. Translationese and its dialects. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*.

Kunz, Kerstin & Ekaterina Lapshinova-Koltunski. 2014. Cohesive conjunctions in English and German: Systemic contrasts and textual differences. In Caroline Gentens, Ditte Kimps & Lieven Vandelanotte (eds.), *Advances in corpus compilation and corpus applications*. Rodopi.

Kurokawa, David, Cyril Goutte & Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of mt-summit xii*.

König, Ekkehard & Volker Gast. 2007. *Understanding English-German contrasts* (Grundlagen der Anglistik und Amerikanistik). 3rd, extended edition. Berlin: Erich Schmidt Verlag.

Lapshinova-Koltunski, Ekaterina. 2013. VARTRA: A comparable corpus for analysis of translation variation. In *Proceedings of the sixth workshop on building and using comparable corpora*, 77–86. Sofia: Association for Computational Linguistics.

Laviosa, Sara. 1996. *The english comparable corpus (ECC): A resource and a methodology for the empirical study of translation*. Manchester, UK: UMIST PhD thesis.

Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. In Sara Laviosa (ed.), *The corpus-based approach: A new paradigm in translation studies*, vol. 4 (META XLIII), 474–480.

Laviosa, Sara. 2002. *Corpus-based translation studies: Theory, findings, applications*. Amsterdam: Rodopi.

Lembersky, Gennadi, Noam Ordan & Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics* 38(4). 799–825.

Matthiessen, Christian. 2001. The environments of translation. In Erich Steiner & Colin Yallop (eds.), *Exploring translation and multilingual text production: Beyond content*, 41–124. Berlin: de Gruyter.

Mauranen, Anna. 2000. Strange strings in translated language: A study on corpora. In Maeve Olohan (ed.), *Research models in translation studies: Intercultural faultlines: textual and cognitive aspects*, 119–141. Manchaster: St. Jerome Publishing.

Neumann, Stella. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin: de Gruyter.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics (acl)*, 311–318.

Popovic, Maja. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *Prague Bull. Math. Linguistics* 96. 59–68.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*. Manchester, UK.

Steiner, Erich. 2004. *Translated texts: Properties, variants, evaluations*. Frankfurt am Main: Peter Lang.

Steiner, Erich. 2012. A characterization of the resource based on shallow statistics. In Stella Neumann Silvia Hansen-Schirra & Erich Steiner (eds.), *Cross-linguistic corpora for the study of translations: Insights from the language pair English-German*. Berlin: de Gruyter.

Teich, Elke. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*. Berlin: de Gruyter.

Toury, Gideon. 1995. *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.

White, John S. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and further approaches. In *Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas*, 193–205.