Chapter 6

Cohesion and translation variation: Corpus-based analysis of translation varieties

Ekaterina Lapshinova-Koltunski
Saarland University

In this study, we analyse cohesion in human and machine translations that we call 'translation varieties' as defined by Lapshinova-Koltunski (2017) – translation types differing in the translation methods involved. We expect variation in the distribution of different cohesive devices which occur in translations. Variation in translation can be caused by different factors, e.g. by systemic contrasts or ambiguities in both source and target languages. It is known that variation in English-to-German translations depends on devices of cohesion involved. We extract quantitative evidence for cohesive devices from a corpus and analyse them with descriptive techniques to see where the differences lie. We include not only English-German translation into our analyses, but also also English and German non-translated texts, representing the source and the target language. Similarities and differences between translated and non-translated texts could provide us with the information on the original of this variation, which might be caused by translationese features.

1 Introduction

This contribution is aimed at the analysis of cohesion in multilingual texts, focussing on variation of cohesive features influenced by different dimensions, i.e. text production type (original vs. translation), translation method involved (manual vs. automatic), as well as systemic contrasts between source and target languages. We know from various studies that translations differ from originals, if various linguistic properties are taken into account (Baker 1995; Teich 2003;

Ekaterina Lapshinova-Koltunski. 2017. Cohesion and translation variation: Corpus-based analysis of translation varieties. In Katrin Menzel, Ekaterina Lapshinova-Koltunski & Kerstin Kunz (eds.), *New perspectives on cohesion and coherence*, 95–118. Berlin: Language Science Press. DOI:10.5281/zenodo.814468

Hansen-Schirra, Neumann & Steiner 2012: and others). These properties of translations distinguishing them from non-translated texts are called *translationese*[1]. In our own studies, i.e. Lapshinova-Koltunski (2015b) and Lapshinova-Koltunski (2015a), we have shown that translations, regardless of the method they were produced with, are different from their source texts and from the comparable originals in the target language. In the latter work (Lapshinova-Koltunski 2015a), we used a set of cohesive features and explorative statistical techniques (automatic classification and correspondence analysis) to discover these differences. In this work, we are using the same set of features, applying descriptive methods which are appropriate for a detailed analysis, zooming into concrete features, such as reference, conjunctive relations and general nouns, as well as their subtypes. This method is supposed to help us to directly compare the differences that we discovered in our previous analyses, and to possibly find out the reasons for the observed variation. Thus, we explicitly compare the feature values and their frequency changes in German and English non-translated texts, as well as human and machine translations from English into German.

Our previous results have also shown that we are not able to discover considerable differences between human and machine translation (MT) in terms of cohesive features if we look at the entire set of features at once. However, we are not convinced that the quality of machine-translated texts can be comparable to that of human-translated ones. For instance, as shown in the examples (1), (2) and (3) ((2) was translated with *Google translate* and (3) was translated by a human), ambiguities cannot be resolved. And in general, translation of coreference and other cohesive devices is poor.

(1) *Alte Mönchsregel: Wenn deine Augen eine Frau erblicken, schlage sie nieder.*

(2) *Ancient monastic rule: When your eyes behold a woman, beat her down.*

(3) *Ancient monastic rule: When your eyes behold a woman, cast them down.*

Although considerable research aimed at enhancing machine-translated texts with textual properties achieved positive results in the recent years, see Webber et al. (2013), Hardmeier (2014) or Meyer, Hajlaoui & Popescu-Belis (2015), document-wide properties of automatically translated texts in terms of coherence still require improvement, as translation models are induced from standalone pairs of sentences. Moreover, target language models approximate the

[1] The term was invoked by Gellerstam (1986).

target language on the string level only, whereas target texts have properties that go beyond those of their individual sentences and that reveal themselves in the frequency and distribution of more abstract categories. Here we mean a certain type of a pronoun or its function instead of the pronoun itself. A more abstract category of the pronouns *he* and *his* would be PERSONAL HEAD or PERSONAL MODIFIER, and for the pronoun *this* – DEMONSTRATIVE HEAD or DEMONSTRATIVE MODIFIER, depending on the context of its occurrence.

We apply corpus-based methods to analyse frequencies and distributions of such cohesive categories in a multilingual corpus that contains English and German originals, as well as multiple translations into German produced with several methods, including manual and automatic ones. Frequencies of cohesive devices will be automatically extracted from the corpus on the basis of automatic pre-processing with a part-of-speech tagger. We are aware of possible errors caused by erroneous tagger output. However, the decision for automatic identification of categories is justified by the fact that we would like to use the knowledge for machine translation, which requires categories that can be annotated automatically with reasonable accuracy. So, we rely on the accuracies of the state-of-the-art tools at hand. The distributions of these categories will then be analysed in originals and translations, as well as in human and machine translations. We will also pay attention to differences between original English and German, as they will serve as a kind of baseline for identifying SHINING THROUGH and NORMALISATION – translationese features resulting from the language contrast between source and target languages.

The obtained information on the differences will be valuable for translation and language contrast studies, and may also find application in multilingual natural language processing (NLP), especially in MT.

2 Theoretical issues and related work

2.1 Cohesion and cohesive devices

COHESION refers to the text-internal relationship of linguistic elements that are overtly linked via lexico-grammatical devices across sentences to be understood as a text, and occurs where the interpretation of some element in the text is dependent on that of another (Halliday & Hasan 1976). Cohesion is related to coherence, whose recognition in a text is more subjective. It involves text- and reader-based features, and refers to illocutionary relations within a discourse. Coherence is the logical flow of interrelated ideas in a text. According to Halliday & Hasan (1976), what distinguishes cohesive relations from other semantic relations is that the lexico-grammatical resources trigger relations that transcend the boundaries of the clause.

The lexico-grammatical devices linking elements in a text and triggering semantic relationships are called cohesive devices. They include personal and demonstrative pronouns and modifiers, substitute forms, elliptical constructions and conjunctions, or lexical devices such as nouns, adjectives and verbs. We will concentrate on two main types of devices: coreference and conjunction, which represent explicit linguistic devices signalling particular conceptual relations to linguistic elements in other text parts (see Halliday & Hasan 1976; Halliday & Matthiessen 2013). These devices are grammar-driven, as most of their items belong to a closed class of functional items.

Coreference and conjunction differ in the conceptual relations that they trigger. Whereas coreference expresses identity to a referent mentioned in another textual part, conjunctions indicate logico-semantic relations between referents, and do not have antecedents, as they do not refer themselves (see Lapshinova-Koltunski & Kunz 2014; Kunz & Lapshinova-Koltunski 2015).

Halliday & Hasan (1976) distinguish three types of coreference: personal, expressed with personal pronouns, possessives and modifiers, as in example (4), demonstrative, expressed by demonstrative pronouns, definite articles, local and temporal adverbs, as well as pronominal adverbs, see example (5), and comparative, expressed by adjectives and adverbs of comparison, as in (6).

(4) *Young men on the roof tops changed their tune; spit and fiddled with the mouthpiece for a while and when [they] put it back in and blew out their cheeks it was just like the light of that day, pure and steady and kind of kind.*

(5) *But no woman ever tried to humiliate him before, to his knowledge, and Fevvers has both tried and succeeded. [This] has set up a conflict between his own hitherto impregnable sense of self-esteem and the lack of esteem with which the woman treats him.*

(6) *Sandy beaches, water sports and activities, evening entertainment and a variety of restaurants make this an ideal base for an active holiday. For a [quieter] and [more relaxing] time or perhaps a walking holiday, go further west…*

As the category of comparative reference is semantically distinct from the first two types (it evokes the relation of similarity or comparison, and not identity, cf.

Halliday & Matthiessen (2004)), we will exclude it from our analysis. Yet, we include another device related to coreference – general nouns. This category is mostly referred to as lexical cohesion, as general nouns are lexical items. However, most of them are cases of abstract anaphora (see Zinsmeister, Dipper & Seiss 2012), or extended reference, and should be, therefore, classified as coreference. In example (7), *this assumption* does not refer to a nominal phrase, but to a clause in the previous sentence. This noun conceptually outlines complex pieces of information, and could also be replaced by the demonstrative pronoun *this*.

(7) *It is only logical to think that if some choice is good, more is better; people who care about having infinite options will benefit from them, and those who do not always just ignore the 273 versions of cereal they have never tried. Yet recent research strongly suggests that, psychologically, [this assumption] is wrong.*

Following Halliday & Hasan (1976), we also distinguish five categories of conjunctive devices classified according to the semantic relations they convey: 1) additive – relation of addition, e.g. *and, in addition, furthermore*; 2) adversative – relation of contrast/alternative, such as *but, by contrast, though*; 3) causal – relation of causality or dependence, such as *because, that is why, therefore*; 4) temporal – temporal relation (*afterwards, at the same time*); and 5) modal – interpersonal and pragmatic relation (*unfortunately, surely, of course*). Most grammars do not include devices of the latter category, which is, however, an important component of a meaningful discourse, as events are connected by speaker's evaluation. Halliday & Hasan (1976) call them 'continuatives'.

2.2 Cohesion in contrastive studies and translation

Cohesion and coherence have been analysed in a number of works on language contrasts dealing with English and German, in which corpus-based methods have become increasingly popular in recent years. However, most multilingual studies are still concerned with individual cohesive devices in particular registers, see Bührig & House (2004) for selected cohesive conjunctions or adverbs in prepared speeches, Zinsmeister, Dipper & Seiss (2012) for abstract anaphora in parliament debates, and Taboada & Gómez-González (2012) for particular coherence relations in a number of different registers. The latter, however, considers also variation in spoken and written language. The authors state that the differences between spoken and written dimensions are more prominent than between languages. Kunz & Lapshinova-Koltunski (2015) and Kunz et al. (2017) also show discrepancy between spoken and written texts, and demonstrate that the distributions of different cohesive devices are register-dependent. The authors show this for a number of cohesive phenomena, analysing structural and functional subtypes of coreference, substitution, discourse connectives, and ellipsis. Their dataset includes several registers, and they are able to identify contrasts and commonalities across languages and registers with respect to the subtypes of all cohesive devices under analysis, showing that these languages differ in the degree of variation between individual registers. Moreover, there is more variation in the realisation of cohesive devices in German than in English. The authors attested the main differences in terms of preferred meaning relations: a preference for explicitly realising logico-semantic relations by conjunctions and a tendency to realise relations of identity by coreference. Interestingly, similar meaning relations are realised by different subtypes of discourse phenomena in different languages and registers.

Cross-lingual contrasts stated on the basis of non-translated data are also of great importance for translation. Kunz et al. (2017) suggest preferred translation strategies on the basis of contrastive interpretations for the results of their quantitative analysis, which show that language contrasts are even more pronounced if we compare languages within each register. These contrasts exist in the features used for creating cohesive relations. Therefore, they suggest that, when translating popular science texts from English into German, translators should use linguistic means expressing cohesive relations more extensively. Overall, they claim that translators should use more explicit devices translating from English into German. For instance, demonstrative pronouns should be used more often instead of personal pronouns: *dies/das* ("this") instead of *es* ("it"). The opposite translation strategies are used when translating from German to English. However, studies of translated language show that translators do not necessarily apply such strategies. Zinsmeister, Dipper & Seiss (2012) demonstrate that translations in general tend to preserve the categories, functions and positions of the source language anaphoras, which results in SHINING THROUGH of the source language preferences (Teich 2003) – in both translation directions. Additionally, due to the tendency to explicate textual relations, translators tend to use more nominal coreference instead of pronominal coreference. EXPLICITATION – the tendency of translations to be more explicit than their sources (Vinay & Darbelnet 1958; Blum-Kulka 1986) – along with *shining through*, belong to the characteristics of translated texts caused by peculiarities of translation process. This translation property forms the focus of studies on the usage of discourse connectives in both manual and automatic translation (see Becher 2011; Bisiada 2014; Meyer & Webber 2013; Li, Carpuat & Nenkova 2014b). Becher (2011) analyses additions (explicitation) and omissions (implicitation) of conjunctive adverbials in business texts, focussing on both English-to-German and German-to-English translations. The author observes more explicitation in the translation direction English-to-German than in the other direction. On the one hand, this is caused by the fact that German has a richer inventory of linguistic triggers for this type of relations (see Becher 2011; Kunz & Lapshinova-Koltunski 2014). But on the other, this is also due to translation properties: they tend towards splitting up information that is presented in one sentence in the source text into two sentences in the target text. This was confirmed by a number of studies (such as Fabricius-Hansen 1999; Doherty 2004; Bisiada 2014). The latter demonstrates that sentence-splitting is a frequent strategy when translating from English into German.

We show that both human and machine translations from English into German differ from their source texts, and also from the comparable German originals, if cohesion and other discourse features are considered (Lapshinova-Koltunski 2015a) . This coincides with one of the features defined within the studies of translationese (see Gellerstam 1986; Baker 1993). According to these studies, translations have their own specific features distinguishing them from the source texts and comparable originals in the target language. One of the features distinguishing them from non-translated texts is levelling out or convergence (Laviosa-Braithwaite 2002) – individual translated texts are more alike than individual non-translated texts. According to Laviosa-Braithwaite (2002), this implies a relatively higher level of homogeneity of translated texts with regard to their own scores in contrast to originals, which would also mean that variation across these texts should be lower than across non-translated ones. As already mentioned above, we believe that translation features are partly effected by the source or the target language involved. Shining through, which was mentioned earlier in this section, is one of these features, and means that we can observe certain features of the source texts in translations. At the same time, we can have an opposite effect, called (over-)normalisation – a tendency to exaggerate features of the target language and to conform to its typical patterns.

2.3 Cohesion in human and machine translations

Differences between human and machine translation in terms of cohesive features have been demonstrated in a number of studies that try to incorporate cohesion-related properties into MT, or use them for MT evaluation.

Li, Carpuat & Nenkova (2014a) show in their experiments that discourse usage may affect machine translation between some language pairs for particular logico-semantic relations. Mascarell et al. (2014) compare translations of German nominal compounds into English, presenting a system that helps to consistently translate coreference via compounds. Guillou (2013) compares lexical consistency (as a part of lexical cohesion) in human and machine translation. Meyer & Webber (2013) analyse explicitation and implicitation of discourse connectives in translation, comparing the occurrence of these phenomena in human and machine translations. Hardmeier (2012), Guillou (2012) and Hardmeier (2014) analyse translation of pronominal anaphora in statistical machine translation, trying to improve performance of their systems.

Most of these studies use human translations as references for evaluating machine ones, whereas direct comparison is carried out in a few cases only. In our own study (Lapshinova-Koltunski 2015a), we compare human and machine translations with each other, and also with comparable source and target texts, analysing a set of cohesive features and their distributions across texts. However, we were not able to show where the differences between human- and machine-translated texts lie, as the observed variation seemed to be more influenced by register than by translation method.

Therefore, in this study, we do not pay attention to the registers that a given text belongs to, and analyse translations applying univariate techniques, assuming that this would allow us to directly observe differences between not only translated and non-translated texts, but also between manual and automatic translations.

3 Methodology

3.1 Research questions

In our analysis we will address several questions related to cohesive devices in English-to-German translations, involving contrastive aspects. These questions are based on the assumptions discussed in relevant works that we described in Section 2 above. We group these questions into three groups: cohesiveness (overall degree of cohesive elements), semantic relations (type of relation used) and variation (variance in data distributions), structuring our analysis (Section 4) according to these.

1. Cohesiveness
    a) How cohesive are the texts in our data?

   b) Are there any differences in the degree of cohesion between translated and non-translated texts, and between different translation methods?

2. Semantic relations
   a) Which semantic relations are preferred over others?
   b) Are these preferences language- or production-type-related?

3. Variation
   a) Is there any influence of language variation onto translations resulting in Shining through/Normalisation?
   b) What are the differences between languages, and between translated and non-translated texts in terms of cohesive devices?

3.2 Data

Our corpus data contains both English-German translations texts and non-translated comparable texts in English and German. English originals (EO, source texts) and German originals (GO, comparable texts in the target language) were extracted from CroCo (Hansen-Schirra, Neumann & Steiner 2012). German translations represent multiple translations of EO, and originate from the VARTRA corpus (Lapshinova-Koltunski 2013). They were produced both manually (human translations) and automatically (machine translations). Human translations were produced by both novice and professional translators. Machine-translated texts were produced with different systems: one trained on a small parallel corpus within a restricted domain, and the other one was trained with a huge amount of unknown data[2].

The whole dataset totals 406 texts which cover seven registers: political essays, fictional texts, instruction manuals, popular-scientific articles, letters to shareholders, prepared political speeches, and tourism leaflets. The decision to include this wide range of registers is justified by the need for heterogeneous data for our experiment (as variation is often register-dependent, see Section 2.2 above). However, in this study, we do not take register variation into account. The total number of words comprises ca. 800.000 tokens. We annotate all texts in the corpus with information on word, lemma, part-of-speech, chunk and sentence boundaries with the help of the TreeTagger tools (Schmid 1994).

[2] See details on the corpus in Lapshinova-Koltunski (2013).

3.3 Feature extraction

As already mentioned in Section 2.1 above, we concentrate on the analysis of two major categories of cohesive devices: coreference and conjunction. We present these categories in Table 1.

Table 1: Features under analysis

device	type	realisation
coreference	pers.pronoun	*he/er, she/sie, they/sie, her/ihr, his/sein, their/ihr, it/es*
	dem.pron	*this/dies/das, that/jenes, this/diese(r/s), that/jene(r/s),* *here/hier, there/da, now/jetzt, then/dann, dagegen, damit*
	gen.nouns	*problem/Problem, situation/Situation, position/Position*
conjunction	additive	*and/und, for example/zum Beispiel*
	adversative	*however/allerdings, in contrast/im Gegensatz*
	causal	*that is why/weshalb, therefore/deswegen*
	temporal	*then/dann, first/erstens*
	modal	*interestingly/interessanterweise, of course/ natürlich*

The first column denotes the category, the second represents their subtypes, and the third illustrates their linguistic realisations (operationalisations) in both English and German. For the extraction of the frequencies of these feature patterns, we use CQP, a corpus query tool (Evert 2005), allowing definition of language patterns in form of regular expressions. These expressions can integrate string, part-of-speech and chunk tags, as well as further constraints, e.g. position in a sentence.

In Table 2, we show examples of the queries for the extraction of personal pronouns (query 1), demonstratives (query 2) and conjunctions (query 3). Queries 1 and 2 contain part-of-speech restrictions only. To further classify them according to their functions (modifier vs. head), we use additional queries with such restrictions as (a) position: before a noun phrase ⇒ modifier vs. no noun phrase following ⇒ head (b) lexical restrictions, especially in case of personal pronouns (*he/him* vs. *his*). Query 3 directly includes lexical restrictions – extracted items should be members of the predefined lists, i.e. additive or adversative conjunctions. An example of the lists is given in Table 3.

Table 2: Examples of queries and extracted examples

	QP query	**example of extracted pattern**
1	[pos=”PP.*”]	*sie, ihr, es...*
2	[pos=”PD.*”]	*dies/das, jenes, diese(r/s)...*
3	[lemma=RE($additive)]	*darüber hinaus, im Weiteren...*

Table 3: Examples of a lists for conjunctions expressing specific relations

<table>
<tr><th></th><th>lexical restrictions</th></tr>
<tr><td>additive</td><td>(
(”(A|a)nd” ”also”)|
(”(A|a)nd” ”yet”)|
(”(F|f)urther”)|
(”(F|f)urthermore”)|
(”(M|m)oreover”)|
(”(I|i)n” ”addition”)|
(”(B|b)esides” ”that”)|
...)</td></tr>
<tr><td>adversative</td><td>(
(”(A|a)lthough”)|
(”(H|h)owever”)|
(”(N|n)evertheless”)|
(”(D|d)espite” ”this”)|
(”(O|o)n” ”the” ”other” ”hand”)|
(”(I|i)nstead”)|
(”(O|o)n” ”the” ”contrary”)|
...)</td></tr>
</table>

For the extraction of general nouns, we use queries containing morpho-syntactic restriction such as the one shown in Table 4. We assume that only general nouns within definite noun phrases are cohesive (cf. *this assumption* in example 7 above). Line 1 in Table 4 allows for a definite article. Alternatively, it can be a demonstrative modifier (defined in line 2). The noun itself is defined with lines 3 and 4, where line 3 specifies the part of speech of the searched element, and line 4 redirects the query to the list of predefined lexical items.

Table 4: Example of a query for general nouns

<table>
<tr><th></th><th>QP query</th><th>explanation</th></tr>
<tr><td>1</td><td>[pos=”ART”&lemma=”d_art”]|</td><td>a definite article OR</td></tr>
<tr><td>2</td><td>[pos=”PDS|PDA.*”]</td><td>a demonstrative modifier</td></tr>
<tr><td>3</td><td>[pos=”NN.*”&</td><td>followed by a noun</td></tr>
<tr><td>4</td><td>lemma=RE($general)]</td><td>whose lemma is a member of predefined list</td></tr>
</table>

We illustrate the list of predefined general nouns with an excerpt in Table 5.

Table 5: An excerpt from a query containing a list of general nouns

<table>
<tr><th>Part of QP query</th></tr>
<tr><td>[pos=”ART”&lemma=”account(.*|s)| action(.*|s)| advantag(e|es)| advice| debate(.*|s)| decision(.*|s)| definition(.*|s)| description(.*|s)| discussion(.*|s)| hypothes(i|e)s| idea(.*|s)| issue(.*|s)| matter(.*|s)| message(.*|s)| method(.*|s)| notion(.*|s)| object(.*|s)| observation(.*|s)| opinion(.*|s)| possibilit(y|ies)| problem(.*|s)| scenario(.*|s)|...</td></tr>
</table>

With the help of such queries, we collect distributional information on frequencies of cohesive devices per text, and also per subcorpus (e.g. representing a translation variety).

3.4 Methods

For our analysis, a number of visualisation and statistical techniques are applied to investigate the distributional characteristics of subcorpora in terms of occurrences of cohesive devices, described in Section 3.3 above. These descriptive techniques will allow us to observe and explore differences between groups of texts and subcorpora under analysis.

We use both parametric and non-parametric tests. The latter, also called distribution-free tests, do not assume that your data follow a specific distribution. We use box plots, which are non-parametric, to see if there are any differences between the subcorpora under analysis in terms of the overall cohesiveness (Section 4.1). They display variation in samples of a statistical population without making any assumptions about the underlying distributed data (e.g. that it is normally distributed). Box plots are median-oriented graphics used to visualise a summary of the distribution underlying a particular sample. They conveniently depict groups of numerical data through their quartiles (the three points that divide the data set into four equal groups, each group comprising a quarter of the data). Box plots have lines extending vertically from the boxes (*whiskers*), which indicate variability outside the upper and lower quartiles. We use notched box plots to reveal if the differences between variables under analysis are significant. According to Chambers et al. (1983), if two boxes' notches overlap in the box plot, then there is no 'strong evidence' that their medians differ. Alternatively, the difference between the medians could be described as statistically significant at the 0.05 level[3].

Turning to the analysis of concrete features, i.e. semantic relations and those of identity, we use bar plots and line charts for visualisation.

Bar plots present grouped data with rectangular bars to show comparisons among categories. The lengths of the bars is proportional to the values that they represent. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. We use bar plots for the visualisation, when not more than two features are involved, e.g. relations of identity vs. logico-semantic ones (Section 4.2), or for the subcategories of the identity relations (Section 4.3.2).

Line plots are used to show frequency of data along a number line. They connect data points of a continuous dependent variable across the levels of an independent variable, illustrating differences across the subcorpora. If the lines are horizontal, there is no difference between the measures compared. Conversely, if there is a slope in the shape of the lines, the subcorpora under analysis show a difference. We use line charts for the analysis of differences based on the distribution of logico-semantic relations, since we have more than two variables at once.

In addition, we apply significance tests to test if the observed differences are significant. For this, we calculate the P-VALUE, which indicates the probability of error or chance in the correlation in our data. The default p-value for the

[3] p-value of 0.05, which is commonly used as a bias for significance measure.

difference to be seen as significant is 0.05. So, if the p-value is lower than or equals to 0.05, the probability that the difference between our variables is due to error or chance is lower than or equals to 0.05, so the difference is significant. For the calculation of p-value, we use Pearson's chi-square test and Student's t-test (Baayen 2008) depending on the number of variables in the test set under analysis.

In the following section, we discuss the findings for each of the questions raised at the beginning of Section 3 above.

4 Analyses

4.1 Overall cohesiveness

We measure the overall cohesiveness of the text in our data as the proportion of cohesive tokens (within cohesive features described in 3.3 above) in the total number of tokens per text. Table 6 gives an overview of the minimum, maximum and median values in the four subcorpora under analysis.

Table 6: Overall cohesiveness of EO, GO and translations

	HU	**MT**	**EO**	**GO**
min.	9.78	10.05	9.86	7.95
max.	28.96	27.94	27.57	24.94
median	16.33	15.85	17.44	17.42

As seen from the table, the English and German originals seem to be similar (if the median values are taken into account) in terms of the overall cohesiveness. This contradicts the findings by (Kunz et al. 2017: 22), who observe more cohesive devices in the German texts than in the English ones in their data. On the one hand, the discrepancy in the results can be explained by the definition of the features under analysis. While we use automatically induced cohesive devices, Kunz et al. (*forthcoming*) operate with manually annotated data. On the other hand, we believe that the cohesiveness values can strongly depend on the texts in a dataset, i.e. cross-lingual cohesiveness in Kunz et al. (*forthcoming*) varies depending on the text registers involved: it is higher for English, if fictional texts and those published on corporate websites are considered. The influence of text variability is also reflected in the minimum and maximum values in the subcorpora, see Table 6, with German originals revealing the lowest ones. The highest maximum value and the lowest minimum value are observed in both translation varieties. However, they also demonstrate a lower proportion of cohesive items in terms of the median value, which means that in general, we observe a reduction of cohesiveness in translation, with machine translation showing the lowest values. This contradicts the phenomenon of explicitation – tendency to spell things out rather than leave them implicit. Assuming that cohesive devices help to explicate coherence relations in a text, we would expect translated texts to be more cohesive than non-translated ones. However, we believe that in this case, we would not need to pay attention to all devices taken together, but to distributions of individual phenomena, e.g. conjunctions expressing logico-semantic relations, or proportion of head vs. modifier functions of pronouns. Moreover, a direct comparison of concrete source texts vs. target texts is also required.

Overall, the median values in Table 6 suggest that the difference between the four subcorpora in our data is not big. We test its significance producing boxplots illustrated in Figure 1. As explained in Section 3.4 above, if two boxes' notches do not overlap, we can observe a significant difference between their medians.

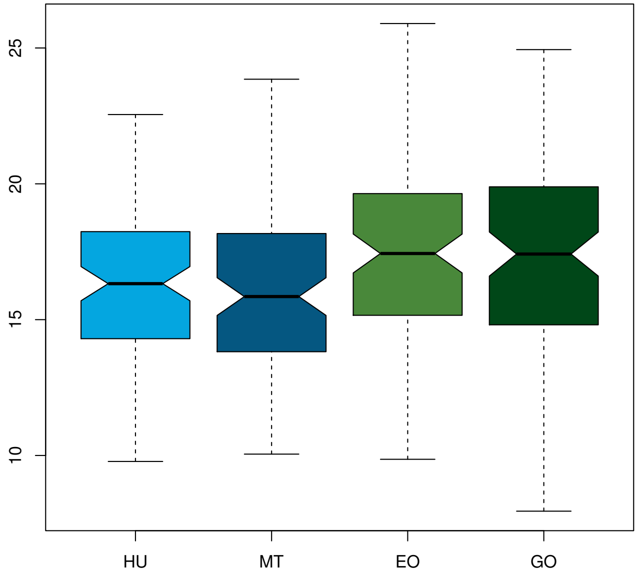

Figure 1: Overall cohesiveness of EO, GO and translations

Analysing notches for the four subcorpora in our data, we see that there is no significant difference between EO and GO, as well as between HT and MT in terms of cohesiveness. Translations (especially machine ones) do differ from non-translated texts, which conforms to the insights from other studies on translationese.

4.2 Semantic relations

In this section, we analyse the distribution of cohesive relations in our data. We start by looking at the distributions for the two main categories: devices signaling identity and devices signaling all types of logico-semantic relations taken together, see Figure 2.

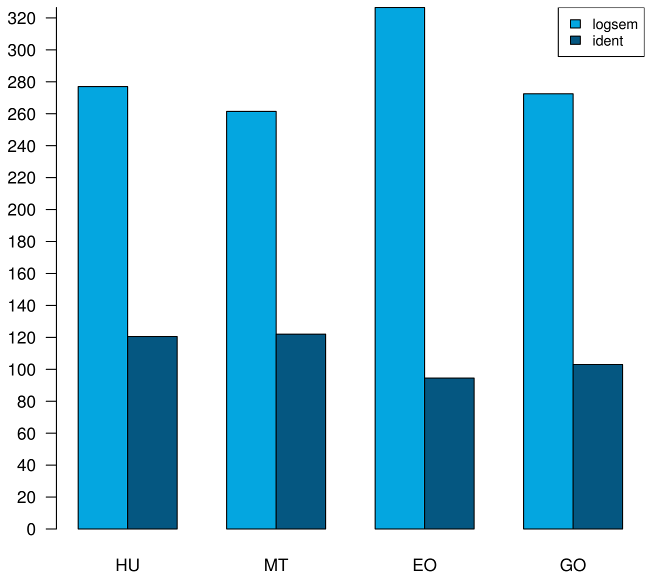

Figure 2: Logico-semantic and identity relations in EO, GO and translations

Figure 2 shows that most of the extracted cohesive data in our corpus is represented by items expressing logico-semantic relations. English texts are characterised by the highest number of logico-semantic devices and the lowest number of linguistic means expressing identity. This contradicts again the results by (Kunz et al. 2017: 25). This discrepancy can be explained by the difference in the definition of conjunctive devices. Ours also include subjuncts which were excluded from the analysis described by (Kunz et al. 2017).

Both translation varieties tend to be similar to German texts. Significance analysis with Pearson's Chi-squared test confirms this observation. The only significant differences ($p < 0.05$) are observed for the pairs EO vs. HT ($p=0.01$) and EO vs. MT ($p=0.004$).

In terms of specific logico-semantic relations, we observe a preference for additive and causal relations in English texts, and for additive and temporal relations for all texts in German (including translations and originals). In this way, our results show that preferences for semantic relations observed in our data are rather language-specific, as translated texts show similarities to comparable non-translated originals in German.

4.3 Variation

In the following, we concentrate on linguistic means expressing cohesion, and their variation across subcorpora under analysis.

4.3.1 Logico-semantic relations

We calculate type-token-ratio (TTR) for cohesive expressions of logico-semantic relations (note that we understand a single occurrence of a conjunctive phrase as a token in this case), see Table 7.

Table 7: Cohesive types expressing logico-semantic relations

	types	tokens	TTR
HU	340	29669	1.15
MT	266	27411	0.97
EO	180	36904	0.49
GO	592	32709	1.81

Although English texts demonstrate the highest number of cohesive items expressing logico-semantic relations, they do not contain many types of conjunctive words. This coincides with general observations on English and German vocabulary, as well as our previous findings (Kunz & Lapshinova-Koltunski 2014), where we also show that the TTR in the German originals exceeds that of the English ones, thus finding a higher degree of variation in the German data. Not surprisingly, both translation varieties reveal a lower degree of variation with a lower TTR. Significance analysis with the help of Student's t-Test shows that the difference between the four subcorpora in terms of TTR is not significant.

Table 8: Ranking of frequent cohesive conjunctions

HT		MT		GO		EO	
7601	und	7939	und	7601	und	12031	as
1225	um	1100	oder	1317	auch	898	because
1162	als	1097	um	1120	als	575	since
995	oder	1004	als	942	oder	459	although
844	wie	879	wie	849	wie	237	but

If we take a look at the five most frequent types (see Table 8), we can see that one main cause for the variance between English and German texts is the high number of occurrences of *as* in EO[4]. Interestingly, the German lists of conjunctions demonstrate discrepancy between non-translated and translated German (*um* in translations, and *auch* in the original German texts). The top three conjunctions in English also explain the preferences for causal relations that we observed in Section 4.2 above.

We must admit that application of fully automatic procedures to extract the data leaves us at the mercy of the tool and the tag set. The TreeTagger does not distinguish between prepositions and subordinating conjunctions which might seriously distort the results concerning conjunction. However, we have to accept these results provided the fact that extractions from all subcorpora under analysis were performed automatically.

4.3.2 Identity via coreference

For identity relations via pronouns, we compare the distributions of their grammatical functions (as a modifier or a head) in all subcorpora under analysis. Figure 3 illustrates functional preferences for coreference with demonstrative pronouns in English, German, and both translation varieties[5].

German texts show the lowest number of modifiers out of all analysed subcorpora, whereas both translation varieties demonstrate a declining number of

[4] Please note that our list can also contain cases of non-cohesive *as*, since all features are extracted with automatic procedures.

[5] The numbers are given in % normalised per total number of tokens.

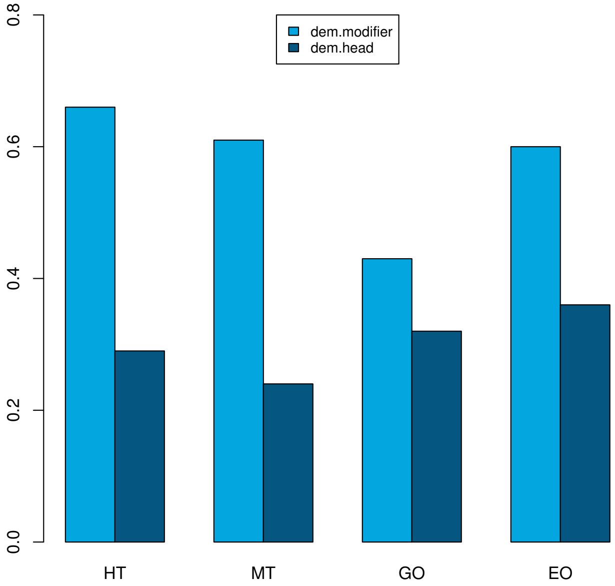

Figure 3: Functional preferences of demonstrative reference

heads[6]. At the same time, we find the highest number of modifiers in translations (with human translation on the top). We assume that this tendency in translation follows from the process of explicitation (see Section 4.1): modifiers that precede a noun or a noun phrase are more explicit means for expressing identity relations than demonstrative pronouns as heads, compare (8) and (9).

(8) *Etwas gerät in Bewegung, und diese Bewegung hält an.* ("Something gets set in motion and this motion continues").

(9) *Etwas gerät in Bewegung, und diese hält an.* ("Something gets set in motion and this continues").

At the same time, it is surprising that translations in our data also demonstrate the highest number of personal heads, as seen in Figure 4.

Analysing variation in the subcorpora with Pearson's Chi-squared test, we find a significant difference between all subcorpora in terms of both personal and

[6] Note that we did not take into account pronominal adverbs, e.g. *darüber*, which also function as heads, as well as definite articles functioning as modifiers.

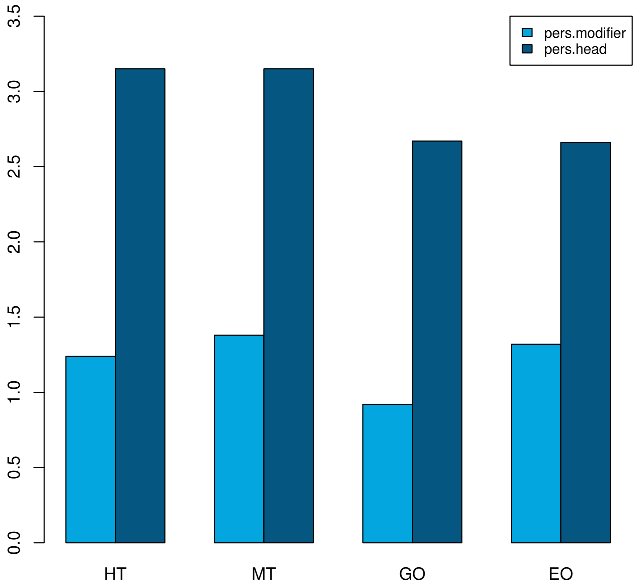

Figure 4: Functional preferences of personal reference

demonstrative reference. The only exception is the distribution of demonstrative modifiers and heads in both translation varieties: human and machine translation apparently do not differ significantly.

In the last step, we compare the type-token-ratio of general nouns. The values for both translation types turn out to be higher than for non-translated subcorpora.

Table 9: General nouns expressing identity relations

	types	tokens	TTR
HT	65	280	23.21
MT	55	191	28.80
GO	99	601	16.47
EO	122	575	21.22

This is surprising, as translations are supposed to have lower TTR than originals (as stated by Hansen-Schirra, Neumann & Steiner 2012). Apart from that, general nouns belong to the most frequent words of the vocabulary. Thus, their higher number in translations might be an indicator of simplification (tendency to simplify the language used in translation).

In Table 10, we present the most frequent general nouns occurring in our data. It is interesting to see that translations share the most frequent general nouns with both source texts (marked with light blue) and comparable texts in the target language (marked with dark blue). We also observe some cases that are common in both English and German texts (marked with light green), as well as words shared by translations only (marked with dark green).

Table 10: Most frequent general nouns

HT	MT	GO	EO
Ziel	Ziel	Frage	system
Weise	Bereich	Ziel	area
Bereich	Problem	Entwicklung	information
Grund	System	Möglichkeit	case
Problem	Ergebnis	Weg	result
System	Frage	Fall	message
Veränderung	Weise	Geschichte	story
Weg	Schritt	Bereich	problem
Ding	Bericht	Prozess	thing
Frage	Punkt	Art	point

Interestingly, human translations share the same number of frequent general nouns with both English and German, whereas machine translations contain more nouns occurring in the English source texts. This is interpreted as a sign of stronger shining through in MT. The word *Weise* shared by both translation varieties is semantically related to *Art* (*Weise* and *Art* are synonyms), one of the most frequent general nouns in German originals texts.

5 Conclusion and discussion

In this paper, we have analysed cohesive properties of multilingual texts that contain both translated and non-translated texts using descriptive techniques. The results show that these properties vary depending on the languages and text production types involved. Languages, even such closely related ones as English and German, have different preferences in the usage of cohesive devices. The observed variation in translations is also influenced by the method involved. Both human and machine translations have constellations of cohesive devices different from those of their underlying originals, and from comparable non-translated texts in the target language. Comparing texts in two translation varieties with original texts in the source or the target language, we found that differences between the two translation varieties are smaller than between translated and original texts. This is not surprising, as parallel data used in the MT development contains human translations. This intensifies levelling out or convergence. We observed this tendency for various features, e.g. for the overall cohesiveness of texts, logico-semantic relations and partly for the relations of identity.

Translations seem to demonstrate explicitation as well, for instance in terms of grammatical functions of cohesive reference via demonstrative pronouns. At the same time, we could not find this for all cohesive devices under analysis taken together. Here, we observed signs of normalisation instead. We could also detect shining through effects, i.e. in terms of general nouns, especially in machine translation.

Overall, our results partly coincide with the observation in our previous analyses: for instance, in our study on shallow features (Lapshinova-Koltunski 2015b), in the one on register-based features (Lapshinova-Koltunski 2017) or the study in which we used discourse-related feature set (Lapshinova-Koltunski 2015a) but applied automatic classification techniques.

At the same time, we realise that there are some limitations of our approach, especially in terms of features under analysis. The frequencies of the cohesive devices were obtained in a completely automatic annotation and query approach. Therefore, on the basis of our findings, we cannot conclude that the processes observed are specific for English and German in general. However, this approach is sufficient for the analysis of differences between the subcorpora at hand, since the features were automatically extracted from all of them.

In the future, it would be interesting to see if the differences between translated and original texts affect perception of the quality of the text as received by humans, for which experiments involving human judgements are required. Moreover, we would like to apply the knowledge on the discrepancies in cohesive devices between human and machine translations, as well as between English and German texts to machine translation, including both MT development and MT evaluation.

Acknowledgments

This paper is based on corpus resources built up in the framework of the VARTRA project[7] which was funded through *Anschubfinanzierung* of Saarland University. We thank our anonymous reviewers for their constructive comments. We also gratefully acknowledge the help of editors in preparing the final version of this article. All remaining errors and misconceptions are our own.

References

Baayen, R. Harald. 2008. *Analyzing linguistic data. A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

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: Benjamins.

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

Becher, Viktor. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts.* Universität Hamburg PhD thesis.

Bisiada, Mario. 2014. Lösen Sie Schachtelsätze möglichst auf: The impact of editorial guidelines on sentence splitting in German business article translations. *Applied Linguistics* 3.

Blum-Kulka, Shoshana. 1986. Shifts of cohesion and coherence in translation. In Juliane House & Shoshana Blum-Kulka (eds.), *Interlingual and intercultural communication*, 17–35. Tübingen: Gunter Narr.

Bührig, Kristin & Juliane House. 2004. Connectivity in translation: Transitions from orality to literacy. In J. House & J. Rehbein (eds.), *Multilingual communication*, 87–114. Amsterdam: Benjamins.

Chambers, John M., William S. Cleveland, Beat Kleiner & Paul A. Tukey. 1983. *Graphical methods for data analysis* (The Wadsworth Statistics/Probability Series). Boston: Duxbury Press.

Doherty, Monica. 2004. Strategy of incremental parsimony. *SPRIKreports* (25).

Evert, Stefan. 2005. *The CQP query language tutorial.* CWB version 2.2.b90. IMS Stuttgart.

[7] http://fedora.clarin-d.uni-saarland.de/vartra/

Fabricius-Hansen, Cathrine. 1999. Information packaging and translation: Aspects of translational sentence splitting (German–English/Norwegian). *Studia Grammatica* 47. 175–214.

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

Guillou, Liane. 2012. Improving pronoun translation for statistical machine translation. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, 1–10.

Guillou, Liane. 2013. Analysing lexical consistency in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, 10–18. Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-3302.

Halliday, Michael A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman Publishing.

Halliday, Michael A. K. & Christian Matthiessen. 2004. *Introduction to functional grammar*. 3rd edition. London: Arnold.

Halliday, Michael A. K. & Christian Matthiessen. 2013. *Halliday's introduction to functional grammar*. London: Routledge.

Hansen-Schirra, Silvia, Stella Neumann & Erich Steiner. 2012. *Cross-linguistic corpora for the study of translations. Insights from the language pair English-German*. Berlin, New York: de Gruyter.

Hardmeier, Christian. 2012. Discourse in statistical machine translation. A survey and a case study. *Discours* 11. Caen.

Hardmeier, Christian. 2014. *Discourse in statistical machine translation*. Uppsala: Acta Universitatis Upsaliensis PhD thesis.

Kunz, Kerstin & Ekaterina Lapshinova-Koltunski. 2014. Cohesive conjunctions in English and German: Systemic contrasts and textual differences. In Vandelanotte Lieven, Kristin Davidse, Caroline Gentens & Ditte Kimps (eds.), *Recent advances in corpus linguistics: Developing and exploiting corpora*, vol. 78 (Language and Computers - Studies in Practical Linguistics), 229–262. Amsterdam/New York: Rodopi.

Kunz, Kerstin & Ekaterina Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies* 14(1). K. Ajmer & H. Hassegard (eds.). 258–288.

Kunz, Kerstin, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski, Katrin Menzel & Erich Steiner. 2017. GECCo – an empirically-based comparison of English-German cohesion. In G. De Sutter, I. Delaere & M.-A. Lefer (eds.), *New ways of analysing translational behaviour in corpus-based translation studies*. TILSM series. Berlin: Mouton de Gruyter.

Lapshinova-Koltunski, Ekaterina. 2013. VARTRA: A comparable corpus for analysis of translation variation. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, 77–86. Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-2510.

Lapshinova-Koltunski, Ekaterina. 2015a. Exploration of inter- and intralingual variation of discourse phenomen. In *Proceedings of EMNLP 2015 Workshop on Discourse in Machine Translation*. Lisbon.

Lapshinova-Koltunski, Ekaterina. 2015b. Variation in translation: Evidence from corpora. In Claudio Fantinuoli & Federico Zanettin (eds.), *New directions in corpus-based translation studies* (Translation and Multilingual Natural Language Processing), 93–113. Berlin: Language Science Press.

Lapshinova-Koltunski, Ekaterina. 2017. Exploratory analysis of dimensions influencing variation in translation: The case of text register and translation method. In G. De Sutter, I. Delaere & M.-A. Lefer (eds.), *New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies*. TILSM series. Berlin: Mouton de Gruyter.

Lapshinova-Koltunski, Ekaterina & Kerstin Kunz. 2014. Conjunctions across languages, registers and modes: Semi-automatic extraction and annotation. In A. Diaz Negrillo & J. Daz-Pérez Francesco (eds.), *Specialisation and variation in language corpora* (Linguistic Insights 179), 77–104. Frankfurt: Peter Lang.

Laviosa-Braithwaite, Sara. 2002. *Corpus-based translation studies, theory, findings, application*. Amsterdam: Rodopi.

Li, Junyi Jessy, Marine Carpuat & Ani Nenkova. 2014a. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 283–288. Baltimore, Maryland: Association for Computational Linguistics. http://www.aclweb.org/anthology/P14-2047.

Li, Junyi Jessy, Marine Carpuat & Ani Nenkova. 2014b. Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classification system. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 577–587. Dublin, Ireland.

Mascarell, Laura, Mark Fishel, Natalia Korchagina & Martin Volk. 2014. Enforcing consistent translation of German compound coreferences. In *Proceedings of KONVENS 2014*. s.n. DOI:http://dx.doi.org/10.5167/uzh-98540

Meyer, Thomas, Najeh Hajlaoui & Andrei Popescu-Belis. 2015. Disambiguating discourse connectives for statistical machine translation. *Audio, Speech, and Language Processing, IEEE/ACM Transactions* 23(7). 1184–1197.

Meyer, Thomas & Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, 19–26. Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-3303.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *International conference on new methods in language processing*, 44–49. Manchester, UK.

Taboada, Maite & María de los Ángeles Gómez-González. 2012. Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences* 6(1–3). 17–41.

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

Vinay, Jean P. & Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais. Méthode de traduction*. Paris: Didier.

Webber, Bonnie, Andrei Popescu-Belis, Katja Markert & Jörg Tiedemann (eds.). 2013. *Proceedings of the Workshop on Discourse in Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-33.

Zinsmeister, Heike, Stefanie Dipper & Melanie Seiss. 2012. Abstract pronominal anaphors and label nouns in German and English: Selected case studies and quantitative investigations. *Translation: Computation, Corpora, Cognition* 2(1). http://www.t-c3.org/index.php/t-c3/article/view/16.