Linguistic profiles of translation manuscripts and edited translations

Tatiana Serbina
RWTH Aachen University
Kármánstraße 17-19
52062 Aachen, Germany
tatiana.serbina@ifaar.rwth-aachen.de

Mario Bisiada
Universitat Pompeu Fabra
C. Roc Boronat, 138
08018 Barcelona, Spain
mario.bisiada@upf.edu

Stella Neumann
RWTH Aachen University
Kármánstraße 17-19
52062 Aachen, Germany
stella.neumann@ifaar.rwth-aachen.de

Abstract

A range of studies have pointed to the importance of considering the influence of editors in studies of translated language. Those studies have concentrated on particular features, which allowed them to study those features in detail, but also prevented them from providing an overall picture of the linguistic properties of the texts in question. This study addresses this issue by conducting a multivariate analysis of unedited and edited translations of English business articles into German. We aim to investigate whether translation manuscripts have a characteristically different distribution of lexico-grammatical features compared to edited translations, and whether editors normalize those features and thus assimilate the translations to non-translated texts. Findings related to individual features are in line with the previously observed phenomena of sentence splitting and passive voice, and a general tendency towards increasing readability. In general, however, no profound effect of editorial intervention could be observed, even though there was a slight tendency of edited translations to be more similar to comparable originals.

1 Introduction

The aim of this study is to assess the role of editors in the translation workflow. This is achieved using the geometric multivariate analysis (GMA) proposed by Diwersy et al. (2014) and Evert and Neumann (2017) to obtain a holistic account of the linguistic properties that characterize translation manuscripts and edited translations. Our pilot study focuses on the first two steps of GMA, namely performing a Principal Component Analysis (PCA) and visually inspecting its results. Specifically, in this paper we address the following questions:

- Do translation manuscripts have a characteristically different distribution of lexico-grammatical features compared to edited translations?
- Do editors normalize the lexico-grammatical features of translation manuscripts, assimilating them to the comparable non-translated texts?

2 Editorial influence in the translation workflow

A number of recent studies analyzed editors' influence on translated texts raising awareness of the part editors play in the translation workflow (Kruger, 2012, 2017; Bisiada, 2017, 2018a, 2019).

Bisiada (2016) studied the phenomenon of sentence splitting, which is often considered a feature of translation that occurs depending on structural conventions in the target languages (Fabricius-Hansen, 1996, 1999; Solfjeld, 2008). He critiques that there seems to be the assumption of an "automatism that seems to assume that translators have little choice in the matter, as the structural principles of the languages involved determine whether sentences are split" (Bisiada, 2016, 354), so that sentence splitting almost necessarily occurs in translations from languages that are considered to prefer a higher informational density to those with a lower one. It is assumed not to occur in the opposite direction, that is, when translating from "low density" language into "high density" languages, because the latter have the structural resources to present information in a compact way.

In his study, however, Bisiada (2016, 374) observes a notable amount of sentence splitting in translations from English to German, thus providing evidence to suggest that "sentence splitting is

an explicitating strategy in translated language in general rather than a process that is triggered only in specific translation directions". As he finds that a significant amount of sentence splitting is attributable to editors, he argues that explicitation as a translation strategy cannot account for the observed frequency of sentence splitting and points to a possible attempt by editors to increase readability (Bisiada, 2016, 371–374).

This is evidenced by further research into the corpus: a study of nominalizations finds that half of them "consist of extensive changes that lead to a complete reformulation of the sentence in question", so that "translators may thus be affected to a greater extent by the academic nature of the source texts, which conventionally favours a nominal style in German, while the editors in this case incorporate popularising strategies" (Bisiada, 2018a, 46–47).

Those findings are corroborated by a study of grammatical metaphorization on the same corpus, which finds that the main influence editors exert is that of turning nominal constructions in the translation back into verbal ones (Bisiada, 2018b,c) as well as turning passive constructions back to active ones (Bisiada, 2019). Both studies suggest that it is through editorial influence that the published translation receives its notable usage frequency of nominal and passive forms.

Kruger (2017) reports on an ongoing study of 208 English non-translated texts, in both unedited and edited form, from the registers "academic, instructional, popular writing and reportage" (Kruger, 2017, 125). To study the influence of editors on the text, she uses a range of operationalizations as proposed by Kruger (2012); Kruger and van Rooy (2012). Her findings are that editors "prefer explicit, non-redundant, analytical constructions, which also tend to be associated with formal writing", most evidently so "in the popular register, where editors' conventionalising impulses override the register preference for more informal usage" (Kruger, 2017, 146). She further reports "support for the hypothesis that editors demonstrate a tendency towards conventionalization or normalization", though they "reduce conventional lexical patterning in the most-frequent range of trigrams" (Kruger, 2017, 146). The study also supports the view that editors simplify the texts.

Bisiada (2017) has further pursued this idea by studying how translation and editing are different activities as regards explicitation, normalization and simplification. The aim of the study was to address the claim that translation universals are really "mediation universals" (Chesterman, 2004; Ulrych and Murphy, 2008) and that editing and translating are thus comparable linguistic activities. This notion was contested by Kruger, who finds a "consistent difference between the translated and edited subcorpus" (Kruger, 2012, 380) in her data.

Bisiada (2017, 268) finds two significant differences: one is between (manuscript and edited) translations and (edited) non-translated texts in the "universal" of normalization/conservatism, the second is that, in terms of simplification, manuscript translations differ from edited texts (translations and non-translations). Bisiada (2017, 268) argues that "editors' influence has been strongest in this respect" and suggests that this may be because simplification is operationalized mainly by quantitative features, which also attract "speed editing" (Bisaillon, 2007, 306).

In terms of a comparison to Kruger (2017), the editorial tendencies towards simplification is corroborated, but Bisiada (2017) finds no reduction of conventionalized lexical patterns in the form of trigrams in translated German; the translations are more conventional than non-translated texts, both before and after editing. This, however, may be due to language differences, corpus composition and also the fact that Kruger (2017) studied non-translations, i.e. texts written originally in the analyzed language, while Bisiada (2017) examined translations.

Bisiada (2017) concludes that the editing stage seems to have had little effect on the features he measured, but states that this "does not mean that changes to the text are negligible, but rather that editors do not intervene in such a way to make the articles more like the non-translated articles" (Bisiada, 2017, 269). This points to the main limitation of research into linguistic properties based on specific features: even if the study takes into account a wide range of them, the picture provided by the results is often fragmented. Observed results are usually interpreted in terms of the specific feature that the analysis concentrated on, which hinders a holistic analysis. This is why we believe that a multivariate analysis provides a full and equal picture to study the lexico-grammatical

features of texts.

3 Methodology

While the above studies have picked a range of individual features for analysis, the present study adopts the multivariate methodology as proposed by Evert and Neumann (2017), whose aim is to study systematic properties of text which, they argue, are not observable on the basis of individual features: "the use of multivariate techniques appears to be essential for a systematic investigation of translation properties" (Evert and Neumann, 2017, 48). The present study therefore runs such a multivariate analysis technique on the corpus compiled by Bisiada (2018a,b,c) (hereafter: Harvard Business Corpus), which was updated by also including text in boxes appearing next to the main articles. The Harvard Business Corpus consists of articles published in the *Harvard Business Manager*, a German sister publication to the *Harvard Business Review*. The articles are translations of articles published in the American edition. The corpus also contains translation manuscripts, which we define as translated texts that were sent by the translation company to the publisher. At least nine different translators have translated the texts at the translation company (in some cases the translator's details were not specified), and six different editors have worked on the texts at the *Harvard Business Manager*.

The articles present findings of scientific studies in an accessible form, geared to managers and business leaders, and thus resemble what is elsewhere known as a popular-scientific format. Others give advice on how to become a better leader or how to manage a company or staff. The magazine sends out specific instructions to its translators where the editors ask them to avoid the nominal style, jargons, the passive and impersonal language use. They are also instructed to dissolve nested sentences (Bisiada, 2016, 356). As these are instructions given to translators, it seems plausible to assume that editors will work with them to hand and use them as their editorial guidelines.

For the present study, this collection of translation manuscripts and edited translations was complemented by a part of the CroCo corpus (Hansen-Schirra et al., 2012). More specifically, in addition to the German translations (edited and non-edited) of business articles (BUSINESS), our data sample includes the published German translations belonging to the registers of letters to shareholders (SHARE) and popular-scientific texts (POPSCI), as well as the German originals from the same registers. Moreover, to counterbalance the size of the data sample consisting of German originals, two additional registers were added, namely the registers of political essays (ESSAY) and prepared speeches (SPEECH). The texts from SHARE and POPSCI were added due to their similarity to the BUSINESS register: letters to the shareholders refer to the performance of the company and the actions of the management, their aim being both to inform and to convince the shareholders. Similar to the business articles, the German translations from POPSCI are mostly articles published in the popular-scientific magazines. Unfortunately, due to the difficulties of finding comparable translations in the opposite translation direction, the sub-corpus of German originals contains popular-scientific book extracts. The aim is to present the scientific findings to the readers in a comprehensible way (Neumann and Hansen-Schirra, 2012). Table 1 summarizes the data used for the present study. The entire data sample consists of 137 texts.

The meta data contains four distinct categories for corpora, namely two different translation versions from the Harvard Business Corpus (Trans – translation manuscripts, Publ – published translations) as well as originals and translations from the CroCo corpus (GO – German originals, GTrans – German translations), and five categories for registers, namely BUSINESS, SHARE, POPSCI, SPEECH and ESSAY.

All texts from Harvard Business Corpus were POS tagged with the STTS tagset (Schiller et al., 1999) using the TreeTagger (Schmid, 1994). The texts from the CroCo corpus that we drew on were tagged using the TnT tagger (Brants, 2000). Based on the previous work on GMA (Evert and Neumann, 2017), the study is based on a set of lexico-grammatical features that were originally defined for the study of register variation (Neumann, 2013). We argue that together the features result in a linguistic profile of the analyzed texts. The process of feature extraction and quantification of every feature per text in the data sample was performed with the help of a CQP script (Fest et al., 2019; Neumann and Evert, Forthcoming).

In the next step, the raw frequencies are normalized using the appropriate unit of measurement, such as nominalizations/words or finite verbs/

Corpus	Translation Status	Register	Size in words
Harvard Business Corpus	manuscript translations	Business	112,810
Harvard Business Corpus	published translations	Business	106,958
CroCo	originals	Share, Popsci, Speech, Essay	137,747
CroCo	published translations	Share, Popsci	69,937

Table 1: Overview of the data sample

sentences. The features that were too sparse in the data and features with correlations r higher than 0.7 were removed from further analysis. From each pair of correlated features, the feature which deemed to be linguistically more informative was kept for further analysis. An overview of the remaining 36 features is shown in Appendix A.

Analysis of the data is performed in two steps. First, the feature counts are discussed descriptively to get the first impression of the data distribution in translation manuscripts and edited translations. In the second step, the features are used as an input for PCA – an unsupervised statistical technique that reduces the number of dimensions within the data set (Levshina, 2015).

4 Analysis

Before performing a multivariate analysis of the data, the distribution of individual features is compared descriptively between the two translation versions contained in the Harvard Business Corpus – translation manuscripts and edited translations. Since the data contains a large amount of outliers, the comparison is based on median that is less sensitive to extreme values. An initial analysis of raw counts showed that translation manuscripts are characterized by more words but contain less sentences as well as less verbs in general and finite verbs in particular. Due to the fact that a lot of other variables are dependent on these values, further comparison is performed using normalized values (see Section 3). For the purposes of this comparison, most of the feature counts are represented here as percentages.

While the differences between the normalized counts are very small, some minor contrasts can be detected (see Table 2, which contains only differences above 1 per cent). These are related to the values of coordination/finite verb, past tense/finite verb, passive/finite verb, and PP as theme/sentence – all of which are used more frequently in translation manuscripts – as well as to adverbs as theme/sentence and conjunctions as theme/sentence – which are slightly increased in the edited translations. Moreover, one further minor contrast concerns the feature words/sentence (the median of 18.82 for manuscript translations, 17.39 for edited translations, 19.31 for non-translations). In contrast to the features included in Table 2, the feature words/sentence represents the number of words per sentence, rather than the proportion. Therefore, this feature count was not transformed into percentage. When compared to medians of the non-translations, the values for all of these features, with the exception of coordination/finite verb and words/sentence, are higher in both translation versions (see Appendix B for the corresponding boxplots).

In order to perform PCA based on the analyzed features, some further preliminary data processing steps are required. In accordance with GMA procedure introduced in Diwersy et al. (2014) and Evert and Neumann (2017), visual inspection of plots plays an important role both during data preparation and interpretation of results. Due to different ranges and distributions of features visible in box plots, normalized feature counts are standardized as z-scores. In the next step, to reduce the influence of outliers, we applied the signed logarithmic transformation of z-scores. Visual inspection of the PCA with and without the log-transformation revealed that individual outliers were reduced, while the overall shape of the data stayed similar. Therefore, all further analyses are performed using log-transformed values. In these analyses every text is projected into a multidimensional feature space as a feature vector comprising the log-transformed z-scores of 36 indicators. The Euclidean distances between the feature vectors are assumed to represent meaningful differences between texts in terms of the selected lexico-grammatical features (Evert and Neumann, 2017).

PCA returns a ranked list of latent dimensions

Feature	Manuscript translations	Edited translations	Non-translations
pasttense/S	29.46	27.7	9.56
passive/F	11.13	6.66	7.69
coordination/F	40.49	38.63	44.44
prepinitial/S	17.02	14.96	7.17
advinitial/S	15.75	17.29	6.67
textinitial/S	2.45	3.73	2.2

Table 2: Distribution of individual features in per cent

characterizing the data. In the present study, over a half of squared Euclidean distance information, identified through the proportion of variance R^2, is captured in the first four dimensions. Figure 1 shows a scatterplot matrix of these four PCA dimensions: the y-axis in each of the rows corresponds to dimensions 1–3, whereas the x-axis in each of the columns corresponds to dimensions 2–4. For instance, the top left plot shows dimension 1 on the y-axis and dimension 2 on the x-axis. While PCA is unsupervised (i.e. meta information such as corpora or registers is not part of the statistical analysis), this information is visualized in the scatterplots to facilitate interpretation of the results.

As can be seen in Figure 1, particularly the first dimension foregrounds the register differences. However, the separation of the five registers present in the data is not complete. Looking at the first dimension, we can see that texts from the BUSINESS register are grouped together mostly on the negative side of the y-axis. Several texts from the POPSCI translations and ESSAY originals are also located on this side. SHARE was placed on the positive side of the axis together with some originals, mainly belonging to the registers of ESSAY and SPEECH. Moreover, around 0 we find another mixed group consisting of almost all texts from the POPSCI register as well as some originals from ESSAY and SPEECH. This distribution is also visible in the density plot shown in Figure 2.

Density curves visualize distribution of texts belonging to the specified categories – in this case the five registers represented in the data – along one of the PCA dimensions. The marks at the bottom stand for individual texts (Evert and Neumann, 2017, 57). The density plot also suggests that the business articles appear to be most similar to the popular-scientific texts.

Analysis of feature weights for this PCA dimension is inconclusive. Similar to the discussion of factor loadings in Factor Analysis, only features with weights below or above the arbitrary threshold of -/+0.3 are considered as significantly contributing to the distribution of texts (Levshina, 2015, 362). Other feature weights cannot be analyzed with certainty. As can be seen in Figure 3, the only linguistic feature with the weight below -0.3 is verbs/word, all other feature weights being in the range between -0.3 and 0.3. Figure 1 shows that business articles are grouped together on the negative side of the first PCA dimension. Therefore, we can conclude that the higher proportion of verbs in business articles is one of the factors that is responsible for this distribution.

While the separation of registers is even less clear along dimension 2, it is worth looking at the distribution of texts by the category of corpus. As shown in Figure 4, all four corpus categories appear to be spread along the whole dimension. However, comparing areas with the highest density per category, we may see a certain tendency of the published translations to be closer to the originals.

Figure 5 shows that the two corpora corresponding to edited and non-edited translations have almost the same distribution between -1 and 2 with the highest density around 0 on the x-axis, whereas all the texts from the CroCo corpus are spread more or less evenly along this dimension.

Dimension 4 does not seem to reflect any interesting patterns in terms of register, corpus or translation status.

5 Discussion

From the perspective of individual features, only slight tendencies could be observed, especially when considering the normalized counts. Some of these differences could be directly related to the previous studies of edited translations. Thus, the higher number of words per sentence and the lower number of sentences together with the

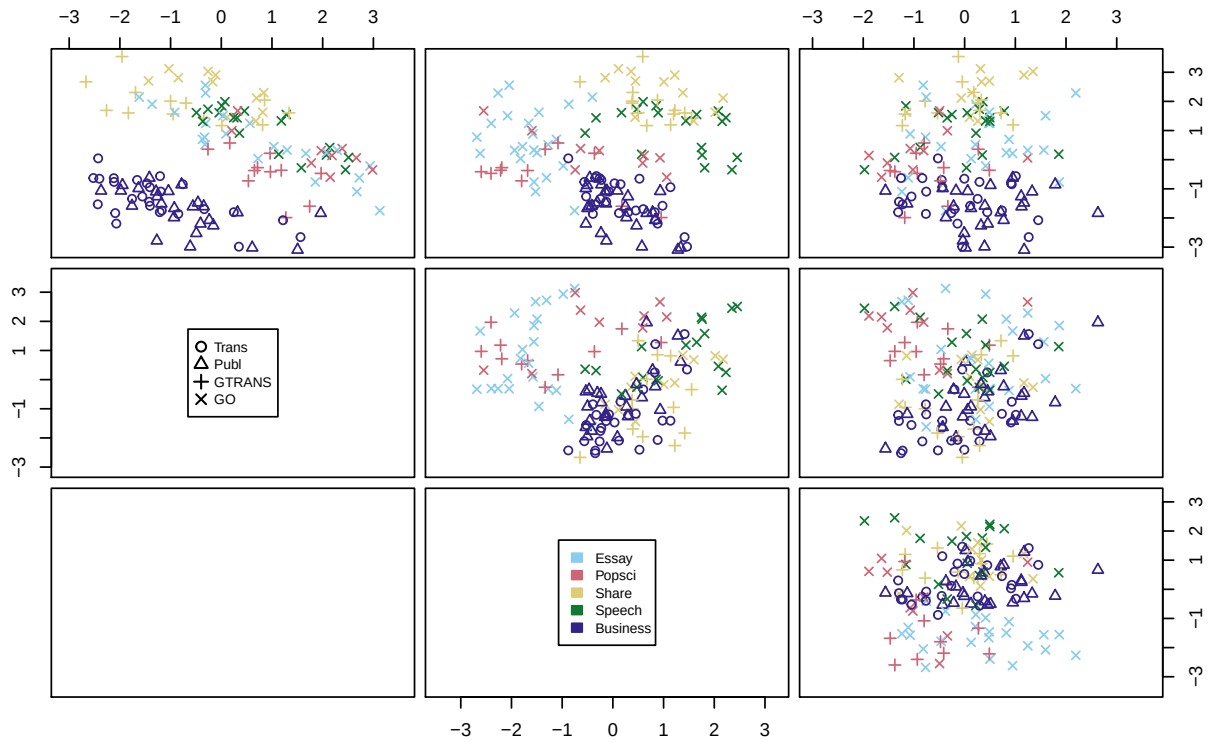

Figure 1: Scatterplot matrix of the first four PCA dimensions

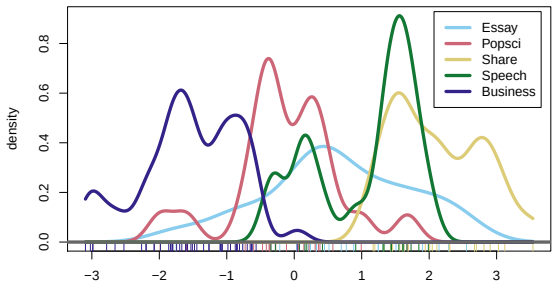

Figure 2: Density plot by register for the first PCA dimension

higher number of coordinating conjunctions attributed to translation manuscripts could be probably explained through sentence splitting (Bisiada, 2016). The difference in terms of the passive voice between the two translation versions within this corpus has been studied by Bisiada (2019). Potential changes by the editors related to the use of past tense and certain elements occurring in the theme position might also be interesting future research questions. The slightly increased numbers for adverbs and conjunctions as theme could indicate a tendency towards introduction of further cohesive devices by the editors – a change that would be in line with the aim of increasing readability of translations. While the comparison of the values to non-translations does not indicate that editors tend to normalize these features, it should be taken into account that the non-translations analyzed in the present study do not contain business articles and are thus not directly comparable to the two translation versions included in Harvard Business Corpus.

From the perspective of a multivariate analysis, we could observe some interesting patterns in the data, even though the identified groups of texts are not clearly separated. Our first research question concerns patterns in the distribution of translation manuscripts and edited translations in terms of their linguistic profiles. Based on the previous research in this area that showed some differences between the two translation versions (see Section 2), we could expect the PCA to separate them into two distinct groups of texts. However, the multivariate analysis did not show a profound effect of editorial intervention. In other words, the combined analysis of the 36 lexico-grammatical features considered in this study suggests that translation manuscripts and edited translations have similar linguistic profiles.

A partial explanation for the differences to the previous research in this area could be a differ-

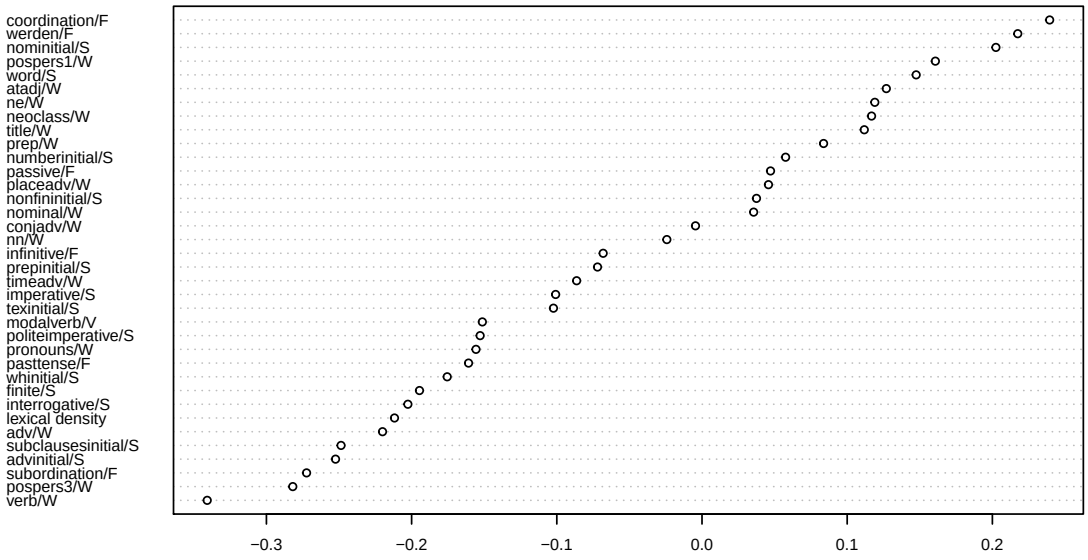

Figure 3: Dot chart of feature weights along the first PCA dimension

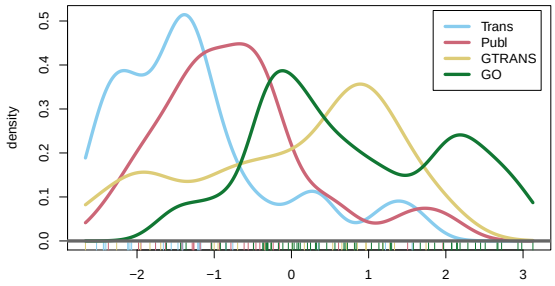

Figure 4: Density plot by corpus for the second PCA dimension

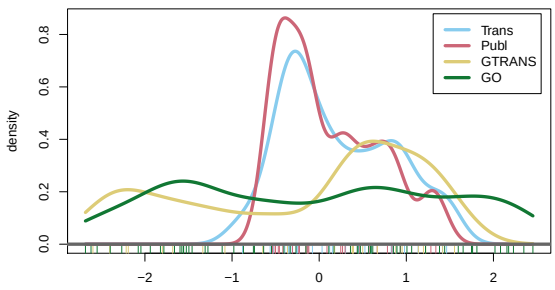

Figure 5: Density plot by corpus for the third PCA dimension

ent type of normalization of feature frequencies as compared to other studies on editorial intervention. As discussed above, a descriptive analysis of normalized feature counts indicated only very minor differences between the two translation versions. For instance, the use of nominalizations, which was reported as one of the differences between edited and non-edited translations (Bisiada, 2018a), is not among the individual features affected by editorial changes when normalized to the number of words per text.

Moreover, and more importantly, our method presents a holistic way of analyzing texts taking into account a large set of linguistic features that together form a linguistic profile. It allows us to generalize on a more global scale than methods focusing on specific features, thus improving our ability to compare text groups in general. That of course does not invalidate existing approaches, as the method applied here cannot detect specific changes that concentrate on a few features and may have a notable effect on the text without changing its overall linguistic profile. The multivariate method we apply in this study of the translation workflow is thus to be seen as complementary to more fine-grained analyses of specific features.

With respect to the second research question

concerned with the translation property of normalization, that is, translated texts being more similar to the comparable originals within the same language (Baker, 1996), we found a slight tendency for some of the edited translations to be closer to the German originals included in our data set, as compared to the translation manuscripts. This means that some of the changes introduced by editors could result in translations being more conventional in the target language, in our case German. The findings should be confirmed using a larger data set. In particular, adding a category of originals comparable to the translation versions included in the Harvard Business Corpus in terms of register, as well as the corresponding English originals could help us explain the unexpected distribution of German translations from the CroCo corpus analyzed for the present paper.

Moreover, the analysis has indicated differences between registers included in our data sample. These contrasts are detected by the most informative first dimension of the PCA. Along this dimension, both translation versions were grouped together as belonging to the same register of business articles. Letters to shareholders, which are comparable to business articles in terms of topic, appear to have very different distributions of analyzed features. In contrast, the popular-scientific register, which is comparable to business articles in terms of aim, seems to have a more similar linguistic profile to the texts taken from the Harvard Business Corpus. One potential explanation could be the fact that our analysis does not contain purely lexical features. It is possible that if individual lexical items were considered as well, then more similarities between business articles and letters to shareholders could be detected. Based on the lexico-grammatical features that are included in the analysis, the results suggest that it is not the topic but rather the aim of texts that is more important for the classification of texts according to register. A follow-up study might consider re-analyzing the business articles as a type of popular-scientific publication.

None of the PCA dimensions has detected differences between originals versus translations within the same language, as was shown, for instance, in Baroni and Bernardini (2006). It is possible that the register effect is so strong that it obscures any effect of translationese.

The present study considers only three sources of variation within the texts, namely translation status (translated vs. non-translated texts), editorial intervention (edited vs. non-edited translations) and register. However, other factors may also play a role. For instance, Figure 5 shows that the CroCo texts are evenly distributed along the third PCA dimension. This might suggest that another source of variation not considered in this study might play a role. It is conceivable that individual variation is responsible for this distribution of texts: taken into account the fact that texts from the CroCo corpus are publications taken from a variety of sources, in contrast to the Harvard Business Corpus, which consists of business articles taken from one magazine, the CroCo texts are likely to contain texts by more individual writers. Unfortunately, both corpora do not contain detailed meta-information, so that it is not possible to include authors/translators/editors as another category that could explain the PCA results.

Following further steps of the GMA procedure (Evert and Neumann, 2017), future research will involve a combination of PCA and a Linear Discriminant Analysis (LDA). This analysis performed on a larger data set involving not only categories considered in the present study but also English originals and German non-translated business articles may lead to finding further meaningful patterns within the data and thus refining the linguistic profiles of translation manuscripts and edited translations.

Acknowledgements

We would like to thank Stefan Evert for developing the R scripts for the GMA procedure and the COMTEX team for modifying the CQP scripts for German. We would also like to thank Florian Frenken for helping us with data pre-processing. Part of the research was funded by the German Research Foundation (DFG) research grant no. NE1822/2-2 and by the Spanish Ministry for Science and Innovation (MICINN), with grant no. PID2019-107971GA-I00.

References

Mona Baker. 1996. Corpus-based translation studies. In Harold Somers, editor, *Terminology, LSP and Translation*, pages 175–186. John Benjamins, Amsterdam.

Marco Baroni and 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.

Jocelyne Bisaillon. 2007. Professional editing strategies used by six editors. *Written Communication*, 24(4):295–322.

Mario Bisiada. 2016. "Lösen Sie Schachtelsätze möglichst auf". *Applied Linguistics*, 37(3):354–376.

Mario Bisiada. 2017. Universals of editing and translation. In Silvia Hansen-Schirra, Oliver Czulo, Sascha Hofmann, and Bernd Meyer, editors, *Empirical Modelling of Translation and Interpreting*, pages 241–275. Language Science Press, Berlin.

Mario Bisiada. 2018a. Editing nominalisations in English–German translation. *The Translator*, 24(1):35–49.

Mario Bisiada. 2018b. The editor's invisibility. *Target*, 30(2):288–309.

Mario Bisiada. 2018c. Translation and editing. *Perspectives: Studies in Translation Theory and Practice*, 26(1):24–38.

Mario Bisiada. 2019. Translated language or edited language? A study of passive constructions in translation manuscripts and their published versions. *Across Languages and Cultures*, 20(1):35–56.

Thorsten Brants. 2000. Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, pages 224–231.

Andrew Chesterman. 2004. Hypotheses about translation universals. In Gyde Hansen, Kirsten Malmkjær, and Daniel Gile, editors, *Claims, Changes and Challenges in Translation Studies*, pages 1–13. John Benjamins, Amsterdam.

Sascha Diwersy, Stefan Evert, and Stella Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, pages 174–204. de Gruyter, Berlin.

Stefan Evert and Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. In Gert de Sutter, Marie-Aude Lefer, and Isabelle Delaere, editors, *Empirical Translation Studies: New Theoretical and Methodological Traditions*, pages 47–80. Mouton de Gruyter, Berlin.

Cathrine Fabricius-Hansen. 1996. Informational density: a problem for translation and translation theory. *Linguistics*, 34(3):521–566.

Cathrine Fabricius-Hansen. 1999. Information packaging and translation. In Monika Doherty, editor, *Sprachspezifische Aspekte der Informationsverteilung*, pages 175–214. Akademie Verlag, Berlin.

Jennifer Fest, Arndt Heilmann, Oliver Hohlfeld, Stella Neumann, Helge Reelfs, Marco Schmitt, and Alina Vogelgesang. 2019. Determining response-generating contexts on microblogging platforms. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*, pages 171–182.

Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner. 2012. *Cross-linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German*. de Gruyter, Berlin.

Haidee Kruger. 2012. A corpus-based study of the mediation effect in translated and edited language. *Target*, 24(2):355–388.

Haidee Kruger. 2017. The effects of editorial intervention: Implications for studies of the features of translated language. In Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, editors, *Empirical Translation Studies: New Methodological and Theoretical Traditions*, pages 113–156. de Gruyter, Berlin.

Haidee Kruger and Bertus van Rooy. 2012. Register and the features of translated language. *Across Languages and Cultures*, 13(1):33–65.

Natalia Levshina. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Benjamins, Amsterdam.

Stella Neumann. 2013. *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. de Gruyter, Berlin.

Stella Neumann and Stefan Evert. Forthcoming. A register variation perspective on varieties of english. In Elena Seoane and Douglas Biber, editors, *Corpus based approaches to register variation*. de Gruyter, Berlin.

Stella Neumann and Silvia Hansen-Schirra. 2012. Corpus methodology and design. In Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner, editors, *Cross-linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German*, pages 21–33. de Gruyter, Berlin.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. *Guidelines für das Tagging Deutscher Textcorpora mit STTS*. Universität Stuttgart, Universität Stuttgart.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Kare Solfjeld. 2008. Sentence splitting—and strategies to preserve discourse structure in German–Norwegian translations. In Cathrine Fabricius-Hansen and Wiebke Ramm, editors, *Subordination versus Coordination in Sentence and Text: A Cross-Linguistic Perspective*, pages 115–133. John Benjamins, Amsterdam.

Margherita Ulrych and Amanda Murphy. 2008. Descriptive translation studies and the use of corpora: Investigating mediation universals. In Carol Taylor Torsello, Katherine Ackerley, and Erik Castello, editors, *Corpora for University Language Teachers*, pages 141–166. Peter Lang, Frankfurt/M.

Appendix A: List of features

Feature name	Description
word/S	Number of words/number of sentences
lexical density	Number of lexical words/number of words
nn/W	Number of common nouns /number of words
ne/W	Number of proper nouns/number of words
nominal/W	Number of nominalizations/number of words
neoclass/W	Number of neoclassical compounds/number of words
pronouns/W	Number of all pronouns/number of words
pospers1/W	Number of 1st person pronouns/number of words
pospers3/W	Number of 3rd person pronouns/number of words
adv/W	Number of adverbs/number of words
atadj/W	Number of attributative adjectives/number of words
prep/W	Number of prepositions/number of words
finite/S	Number of finite verbs/number of sentences
pasttense/F	Number of past tense verbs/number of finite verbs
werden/F	Number of instances of the modal verb *werden* (future)/number of finite verbs
modalverb/V	Number of modal verbs/number of verbs
verb/W	Number of all verbs/number of all words
infinitive/F	Number of infinitives with *zu*/number of finite verbs
passive/F	Number of instances of passive voice/number of finite verbs
coordination/F	Number of coordinating conjunctions/number of finite verbs
subordination/F	Number of subordinating conjunctions/number of finite verbs
interrogative/S	Number of instances of interrogative mood/number of sentences
imperative/S	Number of instances of imperative mood/number of sentences
politeimperative/S	Number of polite imperatives/number of sentences
title/W	Number of titles/number of words
placeadv/W	Number of adverbs of place/number of words
timeadv/W	Number of adverbs of time/number of words
conjadv/W	Number of conjunctive adverbs/number of words
nominitial/S	Number of nominal elements in theme position/number of sentences
numberinitial/S	Number of numbers in theme position/number of sentences
prepinitial/S	Number of prepositions in theme position/number of sentences
advinitial/S	Number of adverbs in theme position/number of sentences
textinitial/S	Number of conjunctions in theme position/number of sentences
whinitial/S	Number of *wh*-elements in theme position/number of sentences
nonfininitial/S	Number of infinitives with *zu* in theme position/number of sentences
subclausesinitial/S	Number of subordinate clauses in theme position/number of sentences

Table 3: List of features

Appendix B: Boxplots

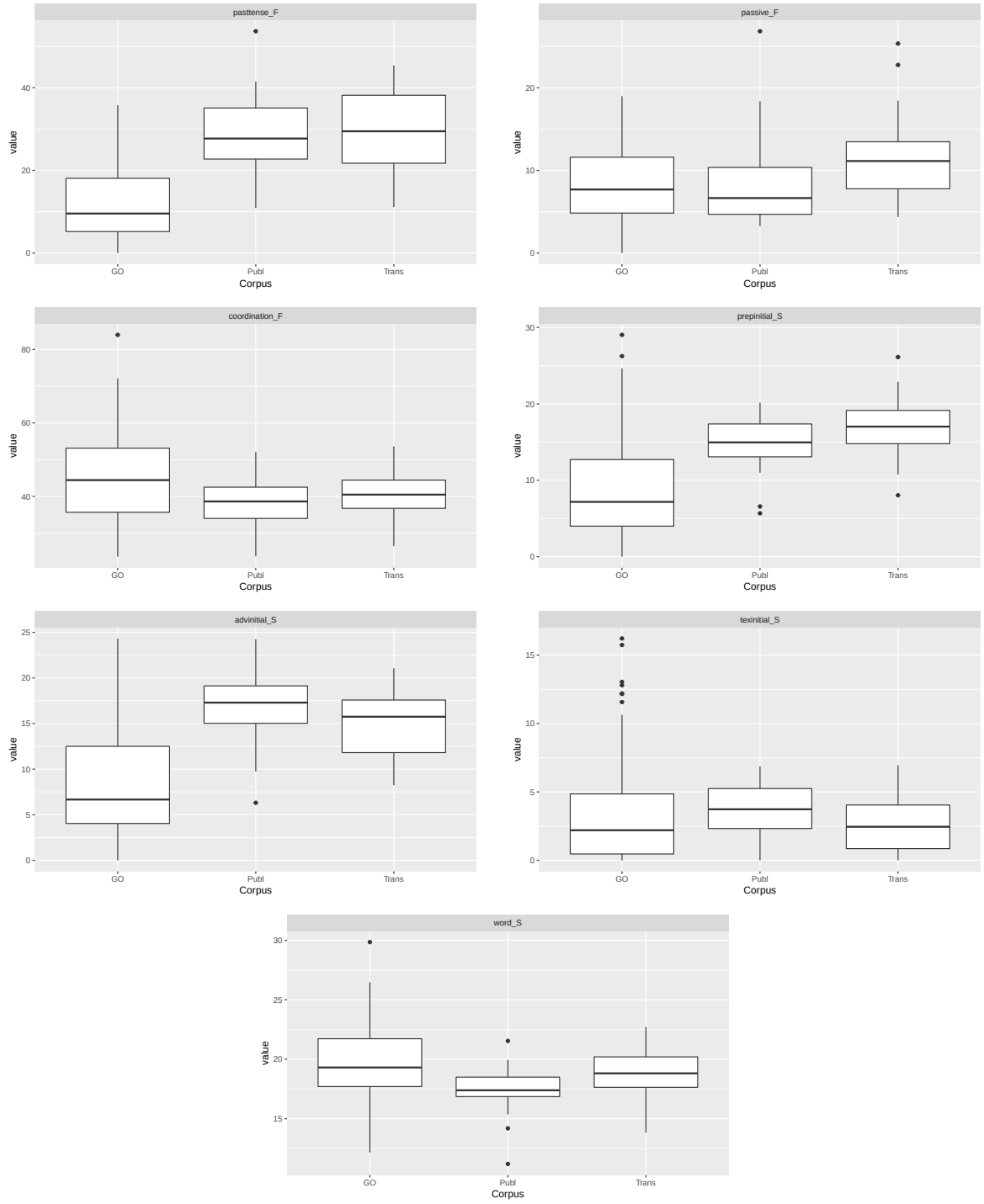

Figure 6: Distribution of selected features across three sub-corpora