The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German

Stefan Evert and Stella Neumann

Abstract

This paper investigates the influence of the source and target language on translations in a selection of 150 pairs of source and target texts from a bidirectional parallel corpus of English and German texts, applying a combination of multivariate analysis, visualization and minimally supervised machine learning. Based on a procedure developed by Diwersy, Evert and Neumann (2014), it investigates the way in which translations differ from comparable original texts depending on the translation direction and other factors. The multivariate approach enables us to detect patterns of feature combinations that cannot be observed in conventional frequency-based analyses, providing new evidence for the validity of interference or shining through in translation. We report a clear shining through effect that is more pronounced for translations from English into German than for the opposite translation direction, pointing towards a prestige effect in this language pair.

1. Introduction

The specific properties that are claimed to distinguish translations from non-translated texts have been the object of research in corpus-based translation studies for almost 30 years. We now have evidence for specific properties of translated versus non-translated text for various language pairs and for various properties (cf. e.g. contributions in Mauranen and Kujamäki 2004; Hansen-Schirra, Neumann, and Steiner 2012 and various individual studies). Many studies however are limited to a restricted set of features: Olohan and Baker (2000), for example, investigate the complementizer *that*, Mauranen (2004) investigates frequencies of lexis, Hansen-Schirra, Neumann and Steiner (2007) analyze cohesive devices. The use of statisti-

cal techniques to draw inferences from the observed patterns in a corpus to the underlying population is still not very well established in translation studies. If a statistical analysis is carried out at all, it is often limited to univariate techniques, e.g. comparing the frequencies of individual linguistic features between translations and originals with Student's t-test or a similar method. A typical example is Neumann (2013), who carries out t-tests comparing translated texts to a reference corpus, focusing on a single linguistic feature at a time. Such univariate methods are suitable for studies examining the effect of a single feature, but they become insufficient when a whole feature catalogue is analyzed.

Systematic properties of text – and that is how translation properties can be characterized – are hardly ever observable on the basis of just a single feature. Most likely, such properties are expressed through a combination of features. Register properties, for instance, may sometimes appear obvious by one individual feature (e.g. imperatives in instruction manuals), but the property of a text serving instructional goals only really emerges if the imperative mood is combined with other features such as short sentences, the use of appropriate terminology, a specific iconic order of clauses in temporal or causal relations, etc. By the same token, individual features hardly ever function in terms of a single property. It is much more likely to assume that one feature contributes to several properties. A high frequency of second person pronouns, for example, can be indicative of reduced social distance and at the same time of the spoken (as opposed to written) medium. Studies that analyze individual features cannot assess correlations between features. Furthermore interactions between different factors that influence the concrete realization of the features are missed. Therefore the use of multivariate techniques appears to be essential for a systematic investigation of translation properties.

Recently, scholars have adopted this approach to profiling translations as compared to non-translated texts. Delaere, De Sutter, and Plevoets (2012) analyze register-related lexical variation as an operationalization of norm-conforming behavior of translators with the help of profile-based correspondence analysis. Contributions in Oakes and Ji (2012) introduce various approaches to the quantitative investigation of translations. Related work by Kruger and van Rooy (2012) draws on analysis of variance (which is not a multivariate technique) to analyze operationalizations of the different translation properties discussed by Baker (1996) across different translated registers in comparison to non-translated texts.

The role of source language interference, one of the features identified as a potential property of translated texts and the main focus of this chapter, was ruled out as a relevant factor on translation by Baker (1993) arguing that it is not related to translation but rather pertains to all kinds of language use where more than one language is involved, such as second language text production. She also argued that a corpus design that collects translations from a wide range of different source languages would level out the influence of the individual source language. However, strictly speaking we would claim that it is methodologically impossible to determine differences between translated and non-translated texts without comparing the realization of a feature in the matching source text: the observed differences might be introduced by other factors than translation effects, e.g. a register divergence between the translations and originally written texts in the same language. Only differences between text pairs aligned at a level appropriate for the respective feature can reliably be claimed to represent properties of *translations* (see Steiner 2012a: 73–75, and on explicitation see Steiner 2012b: 59; for an extensive discussion of aligned pairs of source and target texts see Serbina 2013). Originally described in second language learning as an influence of the L1 on the L2, interference could simply be a general feature of using language in a context where both language systems are activated and trigger choices from both systems in text production (cf. Mauranen 2004). This would mean that, regardless of the specific type of language use (L2 writing and translating into the L1), features from the second activated language system would be likely to interfere with the language in which the text is produced. Interference in this case would not represent a translation-specific phenomenon. While the effect of both types of interference has not yet been sufficiently investigated, we would claim that it is most likely not the same (and also caused by different factors). Interference in second language production involves transfer from the mother tongue into the L2, whereas, at least in the default case of translation into the mother tongue (L1), interference in translation refers to transfer from the L2 into the L1 (see also Steiner 2008 on the directionality of language contact). The comparison between interference in L2 writing and in translation into the L1 is outside of the scope of this paper. Suffice it to say that the specificity of the translation task justifies analyzing interference – or more specifically: shining through – in translation in its own right.

Teich (2003) describes a special case of L2 interference she calls shining through: this property refers to cases where the diverging frequencies of

options existing in both languages are adapted in translated texts to those of the source language, thus resulting in a frequency difference between translations and comparable non-translated texts in the target language. It is this special case of source language-induced divergence of translations that is the focus of this paper. One of the potential factors affecting the extent of L2 interference or shining through could be the diverging prestige of the languages involved (Toury 2012: 314). Toury draws on the sociolinguistic concept to argue that an unequal status of languages and cultures could affect the tolerance of interference. If his claim is right, a difference in prestige between two languages should lead to an asymmetric tendency, in which translations from the more prestigious language into the less prestigious one show more tolerance towards interference than in the opposite translation direction. Mauranen (2004) compares Finnish lexis translated from Russian, a presumably less prestigious culture in the Finnish context, with translations from English, a culture she assesses as more prestigious, and does not find a prestige effect. The claim has been made that the impact of Global English exerts an asymmetric influence on German also by way of translation, to the effect that target culture norms may no longer be maintained which in turn results in convergence with English norms (House 2002: 199–200). Testing this claim, Becher, House and Kranich (2009) report inconclusive evidence that modality is not affected by the contact with English whereas the use of sentence-initial concessive conjunctions seems to converge with English in a diachronic corpus comparison.

On a more general level, Hansen-Schirra and Steiner (2012: 272) describe the relationship between different types of translation-related behavior towards source and target language norms (which in frequency terms can be read as usage preferences) as a continuum ranging from shining through, i.e. orientation towards source language norms, to normalization, orientation towards target language norms.

It is still a matter of debate in translation studies whether such properties are caused by translation-inherent or general factors (cf. Becher 2011 on explicitation). We would claim that the debate could be decided with the help of more comprehensive corpus-based research designs that account for more factors simultaneously: Rather than controlling for register, register variation needs to be assessed as a factor on translation properties. Rather than focusing on individual features at a time, studies should include as many linguistic features as possible and use appropriate statistical techniques to assess these diverse factors and their interaction. Based on the

evidence we now have for instance on the effect of register on variation in translation (Neumann 2013; Delaere 2015), it is obvious that studies concentrating on individual features and controlling for register must inevitably yield contradictory evidence. Finally, rather than excluding the source language, aligned text pairs should be investigated that take into account whether features in a translated text deviate from those in the aligned source text element. The question of the translation inherence of properties can only really be decided on the basis of such improved research designs.

We would claim that the fact that machine learning classifiers are able to distinguish translations from non-translated text with high accuracy provides strong evidence that there are specific traits of translations which need to be explained within the framework of translation studies. In the context of computational approaches, such traits are usually referred to as *translationese*, i.e. some form of distinctive language use in translations. Baroni and Bernardini (2006), for example, report a classification accuracy of 86%, outperforming human annotators. Volansky, Ordan and Wintner (2015) combine the computational approach to translationese with a corpus-linguistic interest in translation properties.[1] They define a set of linguistic features operationalizing translation properties and show that classifiers do not perform equally well across all properties. A finding relevant for our study is that features related to shining through yield the highest accuracy. Despite their success at identifying translated language, these approaches are not geared towards pinpointing factors that might explain the specific make-up of translated texts, or towards detecting hidden structures, e.g. related to differences between translation directions.

In this paper, we use exploratory multivariate techniques to analyze the influence of the source and target language on translations, based on the frequency patterns of different linguistic features in a bidirectional parallel corpus of German and English texts from a range of different registers.[2] To this end, we make the following distinctions. (i) We identify "genuine" shining through of properties of the source language into translations as a general tendency of translators to introduce feature patterns that are typical of the source language into the target texts, quantified in terms of the relative frequencies of comparable lexico-grammatical features. This is distinguished from (ii) text-specific, i.e. individual shining through of idiosyncratic properties of the source texts, reflecting author style, tone, topic domain, etc. In this case, certain linguistic properties of the specific source text are carried over in the translation process. In other words, translators do not adjust their linguistic patterns based on the source language, but

simply translate texts in a relatively literal way. Shining through could also be a side-effect of register divergences between the German and English parts of the corpus. Since this is a special case of individual shining through – the relevant linguistic property being the sub-register of a text rather than e.g. author style – we do not consider this case separately. (iii) These two types of shining through are distinguished from other forms of translationese that cannot be traced back to the respective source language or to individual source texts.

Given the claims about the influence of English on German noted above, we believe that this language pair is a good example for exploring assumptions about source language shining through and more specifically about the impact of translation direction under a hypothesized prestige effect. Our approach is geared towards the type of norm-related translation properties Hansen-Schirra and Steiner (2012) discuss. We will argue that visualization plays a crucial role for understanding the multidimensional structure of the data set.

After a brief introduction of the data and procedure in the next section, we will examine the steps of the multivariate analysis in section 3. Section 4 is devoted to the detailed interpretation of the results of the analysis before these are discussed in section 5 in light of their meaning for translation studies. The paper is rounded off by some concluding remarks and an outlook on future work.

2. Method

2.1. The data

The data used for this study comprise a subset of the CroCo Corpus (Hansen-Schirra, Neumann and Steiner 2012). We discarded the three most extreme registers (novels, instruction manuals and, to a lesser extent, tourism brochures), which accounted for most of the variation in Diwersy, Evert and Neumann (2014) and dominated the unsupervised multivariate analysis, obscuring more subtle, but important patterns such as variation between the remaining registers. We further excluded one text pair as an outlier because the PCA and LDA techniques used by our approach are sensitive to such outliers and give them undue weight in the analysis. In total, we used 298 texts from the five registers political essays (‘essay’), popular-scientific texts (‘popsci’), corporate letters to shareholders

('share'), prepared political speeches ('speech') and websites ('web'). These registers are similar in their focus on factual rather than fictional matters.

The study draws on lexico-grammatical indicators of underlying functions derived in the context of register theory (Neumann 2013). Of the indicators used by Neumann, we included only those which not only exist in both languages but are considered to be comparable, so that original texts and the corresponding translations can meaningfully be compared. We also discarded collinear features, resulting in a final set of 27 indicators which were obtained with a mixture of automatic and manual extraction procedures.[3] A full list of features and their extraction methods is contained in the appendix. All frequency counts are given in relation to an appropriate unit of measurement, e.g. proportion of nouns among all tokens, finites among all sentences, passives among all verbs, imperatives among all sentences, adverbial themes among all themes, contracted forms among all tokens, etc. Additional features are lexical density, the lexical type-token ratio (TTR) and average sentence length (tokens/sentences). To account for the large frequency differences between the various indicators, all values were standardized (z-transformed). The z-transformation also ensures that each feature makes the same overall contribution to the distances between texts described in section 2.2. Every text is thus represented as a feature vector in multidimensional space consisting of the z-scores of 27 lexico-grammatical indicators.

2.2. The approach to multivariate analysis

We adopt the geometric approach of Diwersy, Evert, and Neumann (2014), which makes the assumption that Euclidean distances between feature vectors provide a meaningful measure of the dissimilarity between the corresponding texts and which emphasizes the use of orthogonal projections in order to visualize the geometric configuration of data points in a high-dimensional feature space from different perspectives. This approach has many advantages: First, the position of a text along an orthogonal second dimension does not affect its interpretation with respect to the first dimension. Second, the total variance of the data set – i.e. the average (squared) Euclidean distance between two texts – is the sum of its variances along a set of orthogonal dimensions. We can use the respective proportion of variance (R^2) as a quantitative measure of how much of the geometric configuration is captured by a particular orthogonal projection. Third, the angle between two non-orthogonal axes indicates the amount of overlap

between the information provided by these axes about the data set. If the angle is small, the second axis offers little additional information over the first; if the axes are orthogonal at an angle of 90 degrees, they provide complementary information (cf. the first point made above). Diwersy, Evert, and Neumann (2014) propose the following steps for the multivariate analysis:

1. Apply unsupervised Principal Component Analysis (PCA) to obtain a perspective that captures the overall shape of the data set. PCA yields a ranked list of orthogonal latent dimensions, chosen to maximize the proportion of variance (R^2) preserved by orthogonal projection into the first PCA dimensions.
2. Visualize this perspective with two- and three-dimensional scatterplots, using meta-information such as language, translation status and register to highlight interesting patterns and facilitate the interpretation. The visualization can also reveal methodological problems such as outliers.
3. Introduce a minimal amount of theory-neutral knowledge in order to find a perspective that throws into relief aspects of the geometric configuration which are relevant to the research question. In our case, this leads to a perspective that shows a clear separation of English and German originals even though its R^2 is smaller than for the PCA dimensions.
4. A suitable perspective can be determined automatically using Linear Discriminant Analysis (LDA), a machine learning procedure that maximizes the distance between two (or more) groups while minimizing within-group variability. The LDA discriminant can be used as a dimension for the orthogonal projection and is usually combined with a PCA analysis of the orthogonal complement space for visualization.
5. Validate the LDA model on separate test data to ensure that it has not been overfitted to individual data points. This is usually carried out by cross-validation using Support Vector Machines (SVM) or a similar machine learning classifier. Diwersy, Evert and Neumann (2014: 185) emphasize the importance of this step to avoid circularity and deductive bias. Latent dimensions are identified based on their proven ability to distinguish categories introduced in step 3, rather than on the analyst's subjective interpretation.
6. If necessary, repeat from step 2 in order to improve the analysis. In this paper, we only report the final analysis obtained after several iterations of visualization and interpretation.

7. Develop a linguistic interpretation based on visualizations, quantitative validation, and the (constellations of) feature weights of the LDA discriminant or other latent dimensions. In section 4, the interpretation of feature weights is scrutinized more thoroughly and further developed compared to the discussion in Diwersy, Evert and Neumann (2014).

2.3. Characterization of the approach

In comparison to conventional linguistic approaches, our method does not only support the choice and interpretation of features based on register theory but also gives a global perspective on feature combinations and correlations where, for instance, Neumann (2013) only analyzes the behavior of individual features. Comparing our approach to related work using unsupervised multivariate analysis – in particular Biber's multidimensional analysis (e.g. Biber 1988) – both approaches identify latent dimensions based on feature correlations and thus facilitate the visualization of the high-dimensional distribution of a data set. However, our approach assumes a geometric perspective by focusing on orthogonal projections, in contrast to the Factor Analysis (FA) used by Biber (1988). A key difference is the introduction of weakly supervised information in order to discover more delicate patterns of interest beyond the main dimensions of variation found by an unsupervised analysis (see our discussion in section 3). Our work can also be compared to studies that apply machine learning approaches to translationese (cf. section 1). Our approach goes beyond these by combining machine learning (LDA) with unsupervised multivariate analysis (PCA). We do not operationalize indicators for translationese or translation properties and test their usefulness in machine learning experiments (Volansky, Ordan, and Wintner 2015), but investigate the behavior of indicators derived independently of our translation-related research question (namely in the context of register studies). Finally, unlike studies based purely on machine learning, our analysis emphasizes the importance of visualization, especially since a direct interpretation of feature weights can be misleading (see section 4). Furthermore, visualization allows us to appreciate each data point individually rather than interpreting a summarized and thus inevitably idealized version of the data represented by means.

We use scatterplot matrices, as exemplified by *Figure 1*, to visualize high-dimensional vector spaces. Each panel in such a matrix shows a different two-dimensional perspective on the full space. In *Figure 1*, for example, the top-left, top-center and center panels display three side views of

a three-dimensional cube. However, even trained analysts sometimes find it difficult to discern more complex structures that are not aligned with one or two of the dimensions, and overlapping data points in 2D plots further obscure important patterns. Therefore, we provide 3D animation videos as well as colored versions of some plots in an online supplement to this paper at http://www.stefan-evert.de/PUB/EvertNeumann2017/. The animation for *Figure 1* shows a 3D view of the first three PCA dimensions and rotates through the three side views seen in the scatterplot matrix.

3. Multivariate analyses

Following the procedure described in section 2, we begin with a Principal Component Analysis (PCA) in order to understand the overall geometric shape of the data set. Since it is unsupervised, PCA does not make use of any information on language, translation status or register of the texts, but these attributes can help to highlight structure in a visualization of the data set. *Figure 1* shows the first four PCA dimensions in the form of a scatterplot matrix. Together, they account for R^2=41.9% of the variance of the data set, capturing major aspects of its overall structure. In the plot, German texts are represented by circles, English texts by crosses; originals are shown in black and translations in grey (a color version and animation can be found in the online supplement). The top-left panel, for example, shows the first PCA dimension on the vertical axis and the second dimension on the horizontal axis. The top center panel also shows the first dimension on the vertical axis, but the third PCA dimension on the horizontal axis.

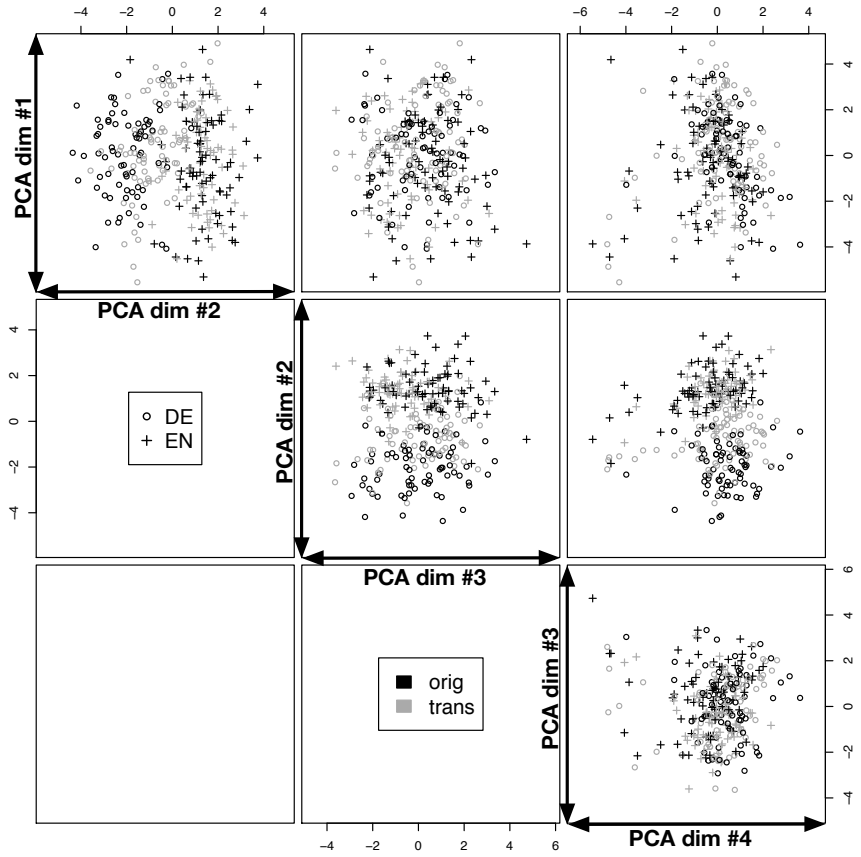


Figure 1. Scatterplot matrix showing the first four PCA dimensions.

The main differences between German and English are captured by the second PCA dimension (horizontal axis of the top-left panel, vertical axis of the two panels in the middle row), which separates the two languages quite well (almost perfectly if translations are excluded). Dimensions 1 (vertical axis of top row) and 3 (horizontal axis of panels in center column) mainly account for register variation, as can be seen from the top-center panel of the register-coded scatterplot matrix in the online supplement. Dimension 4 separates some of the web texts, which appear to be markedly different from the rest of the corpus.

Figure 1 also shows that German translations are shifted towards the English side of the second PCA dimension, while English translations oc-

cupy the same range as English originals. This trend can be seen more clearly by plotting the distribution of texts from the four categories (Germans vs. English, original vs. translation) along this dimension. *Figure 2* shows density curves, which can be thought of as smoothed histograms, with individual data points indicated by the marks at the bottom.

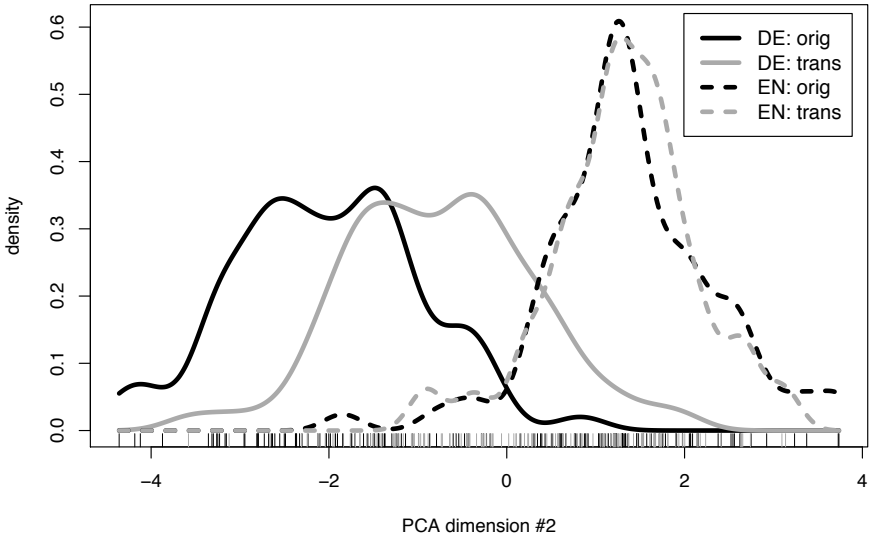


Figure 2. Distribution of texts along the second PCA dimension.

The plot shows an identical distribution for English originals and translations (dashed lines on the right-hand side of the plot), while the German translations are shifted to the right compared to German originals (solid lines on the left-hand side of the plot). While there is more variability among the German texts – shown by a flatter and wider shape of their density curves – German translations and originals follow a very similar distribution, which is merely shifted for the translations. Focusing on the black curves, it is obvious that German and English originals are separated almost perfectly: original texts with a positive coordinate score are mostly English, those with a negative score are mostly German. The central range (roughly from –1 to +0.5) contains very few original texts, but a substantial number of translations into German: apparently, they tend to fall between the originals in both languages.

These observations strongly suggest a shining through effect for translation into the lower-prestige language (German), but not for the opposite translation direction. However, there are a number of issues that need to be taken into consideration before we can draw such a far-ranging conclusion. First, the four-dimensional projection on which our interpretation is based so far accounts for less than half of the total variance of the data (R^2=41.9%). While this is sufficient to give a general idea of the geometric shape of the data set, the remaining 58% – which are entirely invisible in *Figure 1* – may contain further differences between German and English that put the observed shining through pattern in a different light. The characteristic differences between translations and originals that allow machine learning approaches to achieve high classification accuracy must also be hiding in these invisible orthogonal dimensions (especially for English, which shows no evidence for any form of translationese so far).

Second, there is still considerable variability along PCA dimension 2 within each language. In *Figure 2*, many of the German translations fall into a plausible envelope of variation for original German texts, so the observed shift cannot unambiguously be attributed to translation effects. One possible explanation are register divergences between the English and German originals. The German translations might simply represent subregisters that are not covered by the German originals.

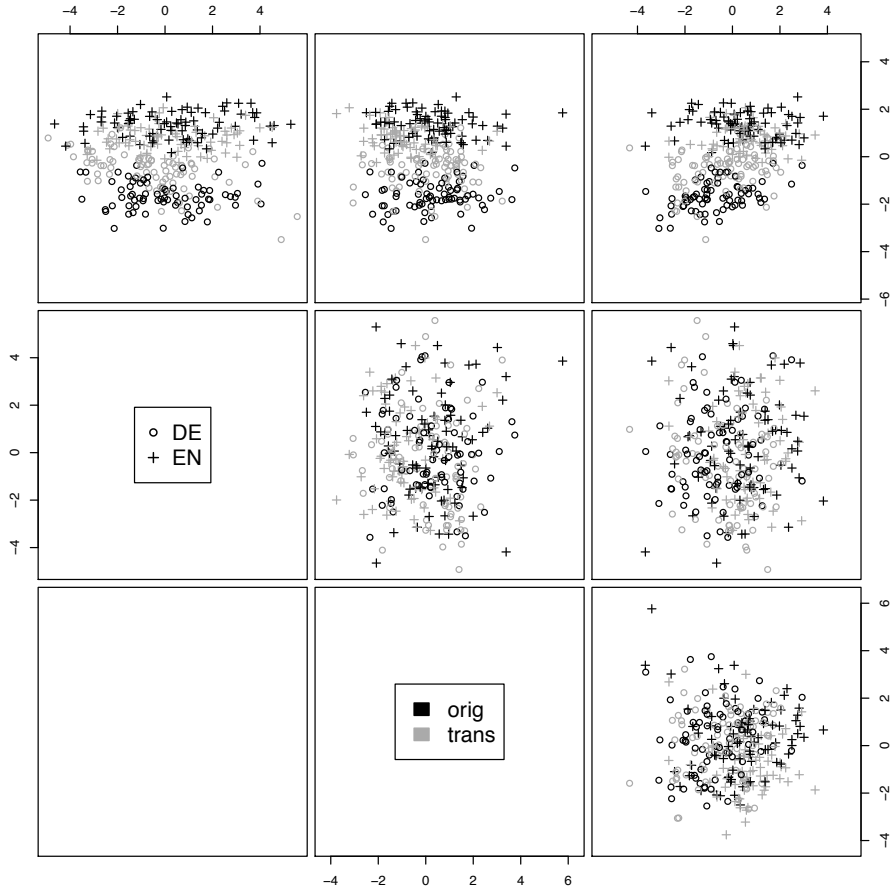
Third, the unsupervised PCA is based on the full data set containing both original and translated texts. It thus captures not only genuine differences between the two languages, but also translation effects, register divergences, etc. If dimension 2 is not based purely on the language contrast between English and German originals, the observed shift cannot directly be interpreted as a shining through effect. Let us clarify this point with a thought experiment: imagine that there is a dimension that captures the language contrast for original texts and a second, completely different dimension that captures a form of translationese introduced by the German translators which is independent of the source language. The PCA might have collapsed these two dimensions into a single axis, so that the shift of German translations from German originals reflects their position on the language-independent translationese dimension rather than actual, i.e. language-dependent, shining through.

In order to focus on the genuine language contrast, we apply supervised Linear Discriminant Analysis (LDA) between the German and English originals, temporarily excluding the translated texts. This procedure is in line with step 3 of Diwersy, Evert and Neumann (2014), adding a minimal

amount of external information; since the learning algorithm is entirely unaware of the translations, there is no risk of biasing the results of the analysis with respect to the shining through hypothesis. The LDA discriminant aims to maximize the separation between German and English originals, while minimizing variability within each language at the same time. Speaking in geometric terms, the discriminant finds a perspective that reveals the most clearly articulated structure, resulting in a clear gap between the German and English originals. It does not account for all differences between the two languages, though, excluding weak tendencies towards higher or lower frequency in favor of characteristic properties separating the languages. As a consequence, the discriminant only captures 6.5% of the total variance of the data, compared to 11.1% for the second PCA dimension. We believe that this approach allows for a better interpretation with respect to the shining through hypothesis: any texts located in the gap between the two groups of originals have properties that are atypical of either language. Forms of translationese which are independent of the source language (type (iii) in Section 1) are very implausible as an explanation for these observations. Note that our focus is not on disproving the existence of (universal) properties of translations, but rather on providing evidence for the existence of genuine shining through in translations. Type (iii) translationese may well exist in addition to shining through, but it does not explain the effect we found.

We can now carry out an orthogonal projection of all texts (both originals and translations) into the one-dimensional focus space defined by the discriminant. For visualization (step 4), the discriminant is extended with PCA dimensions from the orthogonal complement space in order to put the characteristic difference between German and English into perspective. The scatterplot matrix in *Figure 3* shows that the characteristic difference between German and English – i.e. the spread of originals along the vertical axis in the top row – is noticeably smaller than register variation and other effects captured by the PCA dimensions – exemplified most clearly by the wider spread of data points in the panels of the middle row. A scatterplot matrix colored by register and a corresponding 3D animation can be found in the online supplement. Quantitatively, the LDA discriminant accounts for R^2=6.5% of the variance, compared to 15.6%, 8.1% and 7.9% for the first three PCA dimensions.

Figure 3. LDA discriminant for German vs. English originals (vertical axis of top row) with additional PCA dimensions from the orthogonal comple-



ment space.

Focusing on the original texts (black points in *Figure 3*), we see a clear separation of German and English along the LDA discriminant, with only a few "outlier" texts in the gap region. This becomes even clearer in the online supplement where translations are shown in red. Translations in both languages (grey points) extend well into the gap, on the other hand, providing further evidence for a shining through effect, which seems stronger for translation from English into the less prestigious language German. As pointed out above, the LDA discriminant does not capture all differences between the German and English originals, since it focuses on bringing out the most distinctive structure. Dimension 4 (horizontal axis of top-right

panel in *Figure 3*) shows a slight shift between German and English originals: most German originals (black circles) fall in a range from –4 to +2 on this dimensions, whereas most English originals (black crosses) range from –3 to +3. However, variability within each language is much larger than along the LDA axis and the shift is a matter of degree rather than categorization. Like the second PCA dimension in the original analysis, it cannot be used to argue conclusively for or against the shining through effect.

Before taking a closer look at the distribution of translations along the LDA discriminant, we need to validate the supervised LDA (step 5 of Diwersy, Evert and Neumann 2014). We use ten-fold cross-validation to test whether the LDA axis is overfitted to the relative small sample of 149 original texts. In each fold, 90% of the texts are used as training data to compute an LDA discriminant, and the remaining 10% are projected onto this dimension and classified as German or English. With a cross-validated classification accuracy of 97.3% (cf. the confusion matrix in *Table 1*), the distinction between German and English originals is excellent. Discriminant scores of the originals obtained by cross-validation correlate almost perfectly with the scores obtained by the single LDA on all 149 texts carried out above (Pearson correlation *r=.989*). This shows that it is valid to draw conclusions about the language contrast and shining through from the LDA dimension in *Figure 3*.

Table 1. Confusion matrix for cross-classification of originals in LDA.

LDA prediction	true category	
	German	English
German	68	1
English	3	77

For a linguistic interpretation of the LDA discriminant, the feature weights will play a central role. Our findings are only meaningful if these weights are not affected by individual texts in the data set. We can quantify the robustness of feature weights by computing the angle between the full-data LDA and each of the ten LDA discriminants obtained from the cross-validation procedure (see *Table 2*). With an average angle of 9.9 degrees, there is some "wobble" in the LDA dimension, but the general direction of the vector of feature weights remains stable.

Table 2. Angle between LDA discriminant from each cross-validation fold and the full-data discriminant.

	fold 1	**fold 2**	**fold 3**	**fold 4**	**fold 5**
angle	17.6º	14.6º	7.0º	9.7º	4.9º
	fold 6	**fold 7**	**fold 8**	**fold 9**	**fold 10**
angle	9.0º	5.5º	11.3º	10.9º	8.8º

Having confirmed the validity and stability of the LDA discriminant, we can now interpret it as a characteristic difference between English and German originals. Because of the low variability within each language any texts that fall outside these relatively narrow bands have to be considered markedly non-German or non-English. If this holds for translations, these texts exhibit feature patterns that are atypical of the target language, deviating towards typical patterns of the source language: a clear case of shining through. The top-left panel in *Figure 3* already gives a strong indication that this may in fact be the case, Figure 4 displays the distribution of texts along the LDA discriminant in order to confirm this impression.

There is a clear shining through effect for both translation directions, which is more pronounced for translation into German. Note that the two small peaks at the left-hand side of the density curves for German translations are caused by two texts from the web register. Disregarding such individual outliers, the distribution of translations is similar to the distribution of originals in the same language, but shifted by a certain amount towards the source language. The black curves show that German and English originals are separated perfectly by the LDA discriminant (without cross-validation). There is a clearly visible gap at the center that contains hardly any original texts. By contrast, a substantial proportion of the translations (grey curves) are located in this gap and thus are clearly different from originals in either language.

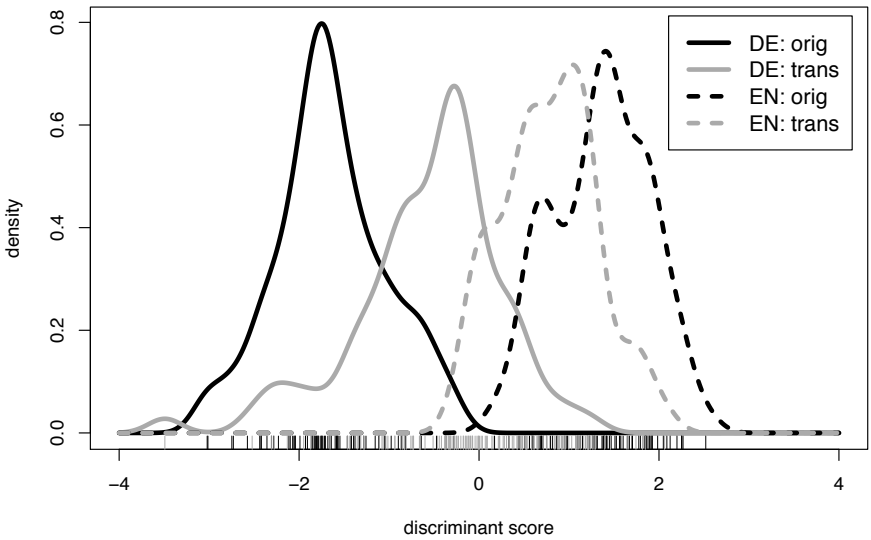


Figure 4. Distribution of texts along the LDA discriminant for German vs. English originals.

These visual impressions now have to be confirmed with a quantitative evaluation (step 6 of Diwersy, Evert, and Neumann 2014). The shift between originals and translations is validated by Student's t-test for independent samples, which shows highly significant shining through in both languages (German: t=9.2378, df=141.54, p=3.4×10^{-16}; English: t = –6.6111, df=145.83, p=6.7×10^{-10}). The effect size (Cohen's d) is 1.5 standard deviations for German, but only 1.1 standard deviations for English, confirming the asymmetry of the effect. Note that the discrepancy between German and English may appear much larger visually, but the higher variability of the German data reduces the relative effect size.

The real test of the shining through hypothesis is whether it is able to account, at least in part, for the marked difference between originals and translations found by supervised machine-learning experiments; i.e., whether we can discriminate between originals and translations based on their LDA scores. Note that the LDA dimension is not overtrained for this purpose because it was determined exclusively based on the originals, without any knowledge about the translated texts. Close inspection of *Figure 4* suggests that LDA scores below –1.1 indicate German originals, scores between –1.1 and +1.3 indicate translations (both into German and into English), and scores above +1.3 indicate English originals. A classifier

using these manually determined thresholds is able to distinguish between originals and translations with 76.8% accuracy, which compares favorably against results reported in the literature (e.g. Baroni and Bernardini 2006), especially considering that those classifiers include translations as supervised training data. In order to exclude the possibility that our thresholds may be overfitted to the data set, we carry out ten-fold cross-validation, using a support vector machine (SVM) with quadratic kernel to select thresholds in each fold. This results in a classification accuracy of 75%–77%, depending on the random split into folds.

We have thus established a clear case of shining through and consequently ruled out other forms of translationese (see Section 1), but there are still two possible explanations for this effect: Rather than showing genuine, i.e. language-specific shining through, the effect could be caused by individual, i.e. text-specific shining through. Note that individual shining through does not necessarily imply that translations are inherently different from originals. The LDA discriminant may have picked up incidental differences between the source texts in the two languages (e.g. because they were sampled from authors with different styles or because of register divergence) that are preserved in the translation and reflected by the shifts in Figure 4.

In order to test whether individual shining through is plausible, we compare the LDA scores of source and target texts in aligned text pairs. If there is individual shining through, we should find a strong correlation between the source and target text. For example, a German text with a very low LDA score should be translated into an English text with a relatively low LDA score and fall into the gap between the originals. A less typically German text with a relatively high LDA score should be translated into an English text with a very high LDA score, overlapping with the English originals. For genuine shining through, this is not the case: a translation tends to exhibit properties of the source language, but its particular LDA score does not depend on the corresponding original text and its LDA score.

Figure 5 and *Figure 6* visualize the correlation between source and target texts. Each point represents a text pair: its horizontal position corresponds to the LDA score of the source text, and its vertical position to the LDA score of the target text. If there is a strong correlation, the points should cluster along a diagonal line. The plots show a difference between the two translation directions, which simply reflects the different ranges on the LDA discriminant occupied by English and German originals (x-axis)

as well as English and German translations (y-axis). However, there is no significant correlation between English originals and their German translations (*Figure 5*; note that the confidence interval includes the possibility of no correlation, r=0), and only a weak, marginally significant correlation for the opposite translation direction (*Figure 6*). Therefore, individual shining through of any kind can be ruled out with high confidence.

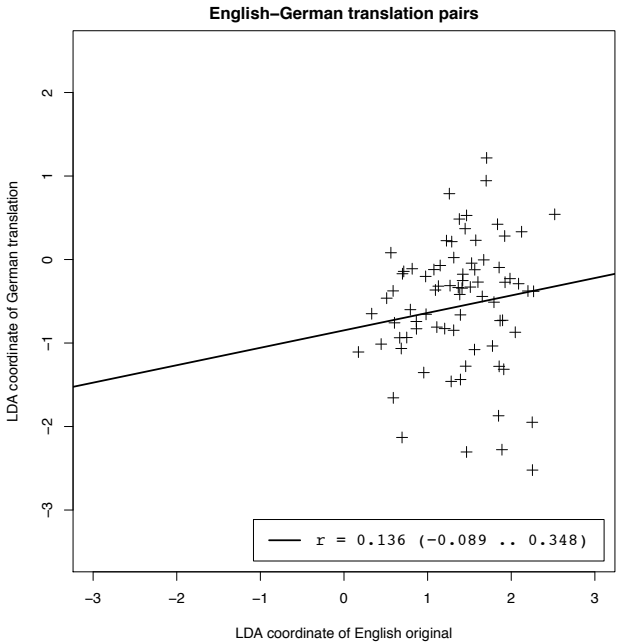


Figure 5. Correlation between LDA score of English originals (x-axis) and their German translations (y-axis), with regression line.

Similar plots for the complement PCA dimensions (not shown here for space reasons) show strong evidence for individual shining through. This does not come as a surprise because the complement PCA dimensions mainly capture register variation, which we expect to be preserved in the translation (e.g., a popular science text should be translated into a text from the corresponding target register rather than an entirely different register). However, the correlation is much stronger than can be explained merely by register effects, in particular along the first complement PCA dimension (dimension 2 in *Figure 3*). We interpret this as evidence for individual shining through of linguistic properties of the source texts that are related to

register and style, but are orthogonal to the contrast between the two languages and thus independent from the genuine shining through effect.

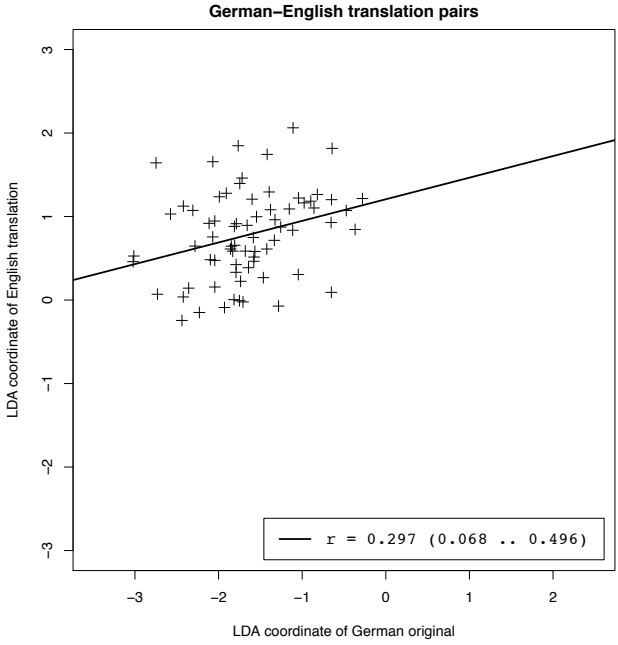


Figure 6. Correlation between the LDA scores of German originals (x-axis) and their English translations (y-axis), with regression line.

Having established a clear type (i) shining through effect in the LDA dimension and verified it with a quantitative evaluation, we can now proceed to the linguistic interpretation and general discussion of our findings.

4. Interpretation of the discriminant

The first step of the linguistic discussion is to determine which lexicogrammatical indicators contribute to the LDA discriminant and hence the observed shining through effect (step 7 of the procedure described in section 2.2). The traditional interpretation of latent dimensions in multivariate studies (e.g. Biber 1988 and related work) focuses on feature weights – as shown in *Figure 7* for our LDA discriminant – and typically applies a cut-

off threshold, disregarding features with absolute weights below the threshold.

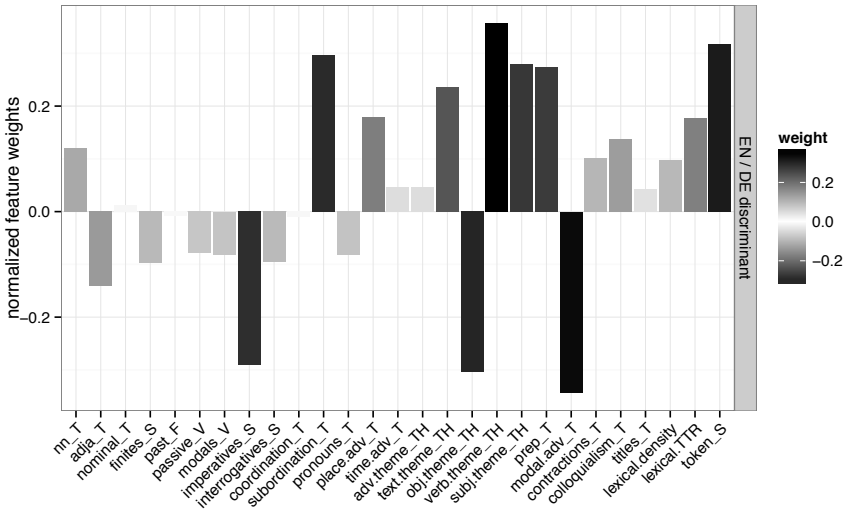


Figure 7. Feature weights contributing to the LDA discriminant (normalized for orthogonal projection).

At face value, positive weights indicate features that are characteristic of English originals (since the English originals are positioned on the positive side of the discriminant axis) and negative weights indicate features that are characteristic of German originals. The traditional interpretation would thus conclude that English originals are characterized by high proportions of textual themes[4], verbal themes, subject themes, subordinations and place adverbs, as well as long sentences (tokens / S) and a high lexical type-token ratio (TTR). German originals are characterized by high proportions of object themes, modal adverbs and imperatives. While such an interpretation may be acceptable for the first few PCA or FA dimensions with their strong correlational patterns, it does not do full justice to the multivariate nature of the analysis because each feature is assessed independently as an indicator of English or German. In our case, this amounts to little more than a traditional univariate language comparison. Consider the boxplots in *Figure 8*, which show the contribution each feature makes to the positions of texts on the LDA axis (i.e. standardized feature values multiplied by the corresponding feature weights), separately for German and English originals. A

feature with a positive contribution pushes texts to the English side of the axis, a feature with a negative contribution pushes them to the German side of the axis. Note that positive contributions correspond to above-average feature values if the feature weight is positive, but to below-average feature values if the weight is negative (indicated by "(–)" in front of the feature name).

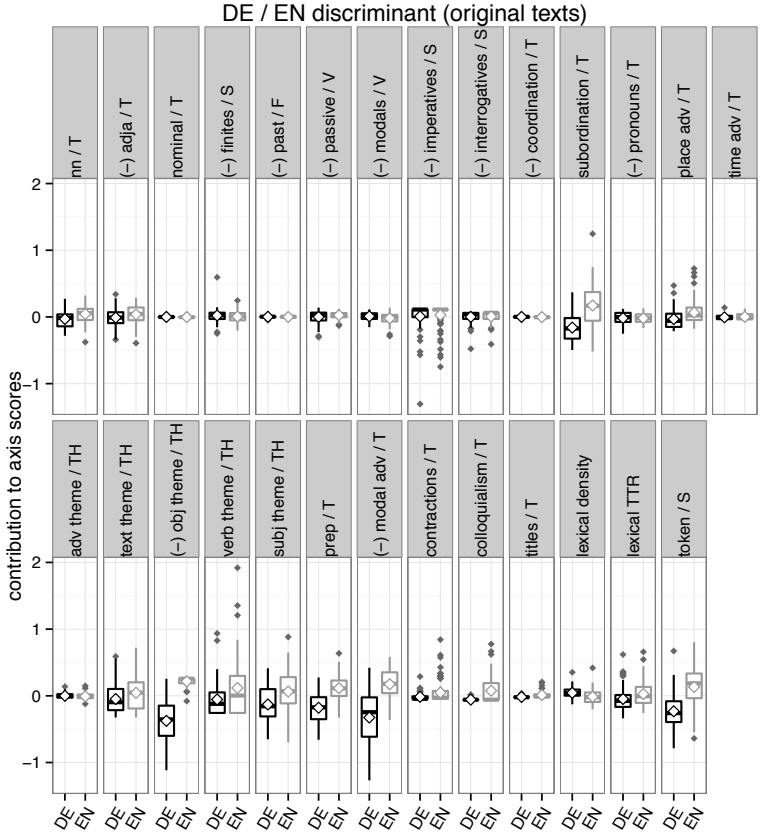


Figure 8. Boxplots showing the contribution of each feature to the position of German and English originals on the LDA discriminant.

Diamond symbols indicate the average contribution of each feature to the positions of German and English originals, respectively. The further its two

diamonds are apart, the more a feature pushes the English and German texts away from each other. However, this does not necessarily mean that the feature improves the discrimination between the two groups: it also adds within-group variability, indicated by boxes and whiskers around the diamonds in the plot. Several features have a very strong effect, including object themes, modal adverbs, subordinations and sentence length (token / S). Other features have a much smaller effect (e.g. textual, verbal and subject themes) or hardly any effect at all (imperatives) despite their large weights. Only one feature (object themes) is highly discriminative by itself, i.e. the boxes for German and English do not overlap: with very few exceptions, only German texts allow themes to be realized as objects. Modal adverbs, prepositions, subordinations and sentence length also contribute well to the language discrimination, while features such as textual, verbal and subject themes seem to add primarily to the within-group variability. Two features (lexical density and modals) even have a counter-intuitive effect: they nudge English originals towards the German side of the discriminant and vice versa. These observations show that an interpretation in terms of feature weights is too simplistic and can be outright misleading in some respects.

As we have already pointed out in section 1, multivariate analysis assumes that features are multi-functional, i.e. they reflect a mixture of several systematic text properties. Ideally, the linguistic interpretation should focus on such underlying properties rather than individual lexico-grammatical indicators, determining which properties account for the language contrast and have thus been found to "shine through" into the target language. The LDA discriminant provides an excellent starting point for this purpose. Since it aims to minimize within-group variability (i.e. among the originals in each language), feature weights are adapted so that the effects of other, irrelevant text properties cancel out. This also explains why some features that have a small effect on the separation of the two groups but large within-group variability (e.g. textual and verbal themes) have nonetheless been included in the discriminant: their main purpose is to help cancel out irrelevant properties.

Due to this complex interplay between features in the underlying structure of the data set a detailed discussion of individual lexico-grammatical indicators will not be attempted here, with one exception. Prompted by the high discriminativity of the single feature object themes, we defined a simplified discriminant based on the four theme-related features with high LDA weights: object, subject, textual and verbal themes. If our assumptions

hold true, this discriminant should represent patterns of theme realization that are characteristic of English and German texts, respectively.

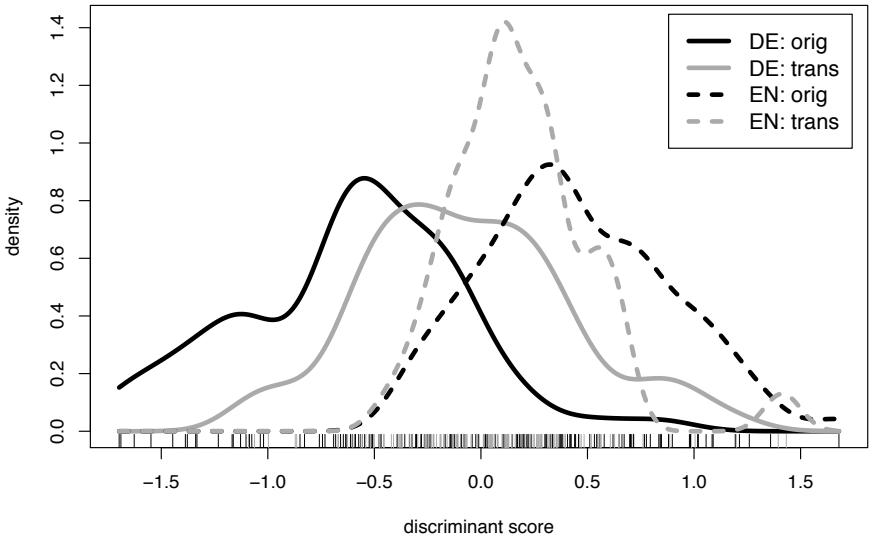


Figure 9. Distribution of texts along a simplified discriminant that represents characteristic patterns in the realization of themes.

Figure 9 displays the distribution of originals and translations along the simplified discriminant. There is still a significant shining through effect, which is stronger for translations from English into German (Cohen *d*=–1.08 for German vs. *d*=+0.65 for English). However, the originals are no longer clearly separated (88% classification accuracy) by the new discriminant. As a result, the translations are not located in a gap that would mark them as clearly distinct from originals in either language. Classification accuracy for translation status is reduced from 76% (for the full discriminant) to 61% (for the theme-related discriminant).

Our conclusion from these observations is that the realization of themes plays an important role in the language contrast between English and German. It is a major factor behind the observed shining through effect. Low classification accuracy shows that this picture is only partial. A full understanding of the LDA discriminant and shining through can only be achieved by exploring the genuinely multivariate patterns of correlations and interac-

tions between the individual features. This is beyond the scope of the present paper, however, and will be addressed in future work.

5. Discussion

Linking our findings to the general discussion in corpus-based translation studies about the character of translation properties, we might ask whether the observations might not support a more generalized claim that shining through is a *universal* feature of translation.[5] The quantitative validation confirmed by the t-test suggests that, at least in the language pair English-German, translated texts can be systematically separated from non-translated texts. This lends additional support to the results obtained by computational studies of translationese (see section 1), now based on more informative lexico-grammatical indicators. Moreover, the interpretation of the visualizations showed that translations in general tend to orient towards the target language, but are still distinctively different in their tendency to accommodate features of the source language.

This finding would support the universals hypothesis. However, the analysis also revealed differences in effect size for the two translation directions thus contradicting this hypothesis because it would require comparable results for both directions. This does not only let the universals hypothesis appear implausible but also makes parallel activation of both language systems and consequently a similar context as in L2 writing (see Section 1) less likely because this scenario, too, would require the effect to be similar in both translation directions. Rather, we have to find additional factors that explain the differential situation for both translation directions. The fact that the effect is stronger for German translations than for English translations can be tentatively interpreted in terms of the differences in prestige discussed by Toury (2012). Note, however, that our study did not test for prestige so that this is just one possible factor that could explain why the translations into German seem to accommodate more characteristics of English as the source language than translations in the opposite direction. The influence of additional factor(s) also provides an argument why the universals hypothesis cannot be upheld. The translator works in too complex a context in which a whole range of factors influences the specific outcome of the translation process. These will interact in various ways depending on their respective strength. At the same time, this finding also further corroborates our initial claim that L2 writing and translation

into the L1 are likely to yield different effects in terms of interference. Incomplete learning of the L2 can be assumed to be an important factor in writing in the foreign language, however, this is a less likely factor for translation – at least into the L1. By the same token, diverging prestige of the languages involved is a plausible explanation for the directionality effect in translation, but cannot be assumed to be a cause of L1 interference in L2 writing.

The analysis in section 3 focused on shining through. Nevertheless, we might also be interested in other properties. Hansen-Schirra and Steiner (2012) describe normalization as being linked to shining through on an assumed norm continuum. Consequently, our study should also reveal this property. Over-normalization, the exaggeration of target language norms, could have become observable in the visualizations if, for instance, the translations had been located on the remote side of the target language originals. While the exact definition of normalization is still a matter of debate (e.g. would not perfect alignment with target language norms be exactly what one would expect?), our study did not yield clear indications of the generalized type of normalization. This could tentatively be interpreted as a reduced importance of generalized normalization, but clearly requires more in-depth analyses in future work. Note that normalization would also be observable, if only part of the translations, say from a register which is particularly prone to covert translation, were located in the expected area. This would be in line with Delaere's (2015) evidence for register-specific target language orientation.

Levelling-out refers to the tendency of translations to converge towards unmarked features at the expense of more marked features that are observable in non-translated texts (Baker 1996). The methodology of this study would also allow us to observe levelling-out, but our experiments did not reveal any notable indications for this assumed property. Cursory examination of individual registers in the PCA dimensions suggests that some registers might display levelling-out, but, again, this has to be relegated to future work. The contentious properties of explicitation and simplification are difficult to investigate with our research design. They could be indirectly included in patterns of shining through, but would probably have to be investigated on the basis of dedicated operationalizations which in turn lead to a risk of circularity in the investigation.

What is the contribution of our study to (corpus-based) translation studies beyond what has already been shown by univariate studies of individual features and registers? Previous studies used differential linguistic features

in order to operationalize properties such as shining through. By contrast, this study focused on features that are actually comparable across the two languages involved. Consequently, the study could have very well produced a quite different outcome showing, for instance, systematic normalization rather than shining through. It provides evidence for the intricate interplay between linguistic features: the overall pattern in the data emerges from a complex combination of features suggesting that findings based on the (cumulative) interpretation of individual features may lead to spurious results that could be counteracted by other features not included in the study. Moreover, our study shows that similar distributional patterns apply across registers, even though we also obtained indications of register-specific behavior in higher PCA dimensions. This will have to be examined in more detail on the basis of a broader coverage of texts and registers in future work.

The results are also of interest from a contrastive linguistics perspective, providing multivariate evidence that the difference between two languages is not only observable in features that only exist in one language but also emerges from the distributional patterns of comparable features.

Against this background, the study also complements claims about the assumed obligatory character of shifts due to contrastive differences. The shining through effect established in our study shows that translators do not necessarily adjust for differences between languages that only consist in usage preferences of comparable features, i.e. differences in their frequencies. In such cases, they do not always adapt the text to match the usage preferences of the target language (see Section 1 and Teich 2003).

While the results of our study look very promising, there are also some clear limitations. As is usual in multivariate analyses, the choice of features and texts heavily impacts the results. This requires eliminating correlated features, computing relative frequencies with respect to appropriate units of measurement as well as avoiding features which cannot be quantified in the same way as the ones discussed here. Especially lexical features, which nevertheless shed light on language variation, can only be included in a quantified, i.e. abstracted form (e.g. in the form of lexical density). More specifically, an analysis of the type presented here requires a large number of lexico-grammatical features, which should be as informative as possible and which need to be extracted in a rather costly procedure. Drawing on automatic analyses makes the extraction of data more efficient but comes at the price of inheriting the inaccuracies of the annotation tools. While each step of the analysis involved great care to ensure reliable data – from select-

ing appropriate tools in Hansen-Schirra, Neumann and Steiner (2012) and establishing comparability of the features in Neumann (2013) to representing the data in our multivariate analysis – the final selection of features is still prone to undue influences. The results reported here need to be read against this background.

Text selection as well as number of texts included in the corpus play an important role. The inherent circularity of sampling texts to be representative of a given set of registers is an important issue that limits the outcome of our analysis. Furthermore, a well-known problem of comparable corpora is the assumption that comparable texts are indeed from comparable registers where, in fact, registers may be slightly diverging. One way of improving this situation is to carry out an annotation of the registers based on external parameters as shown by Delaere, De Sutter and Plevoets (2012). While this does not remedy the potential incomparability between registers in a bilingual corpus, it does facilitate the analysis of the incomparable registers because it helps narrowing down the exact area(s) in which the registers diverge. Furthermore, texts from extreme registers may obscure the behavior of the bulk of the corpus. This was shown to be the case in Diwersy, Evert and Neumann (2014). It is possible to mitigate the effect by eliminating outlier registers, as we did in the work reported here. However, this is not a perfect solution either.

Standardization was claimed to be essential to our approach, so that each feature is given the same weight regardless of the scale of numerical values (see section 2). However, to some extent this may also be a problem because it may increase the influence of individual features. We may be overemphasizing the importance of features that have relatively little variability in language. Passive may, for instance, be relatively frequent across the board; small differences between individual texts are then exaggerated by our approach.

6. Conclusion and outlook

In this paper, we hope to have shown the intricate interplay between languages as well as originals and translations that emerges from the interpretation of latent dimensions of multivariate analysis. More specifically, we reported evidence for a generalized shining through effect of the source language in a corpus of originals and translations from the language pair English-German. To this end, we used a sequence of steps consisting of

PCA, LDA, visualization and cross-validation. The interpretation of the analyses relied heavily on inspecting visualizations that proved to be very informative, throwing light especially on the assumed directionality effect that gives the paper its title. One of our main results is that shining through manifests to differing degrees in the two translation directions, suggesting a tentative interpretation in terms of diverging prestige of the two languages involved.

This hypothesized role of prestige is one of many things that should be examined in more detail in future work. In addition to the aspects already mentioned in the previous sections, this also includes further investigating patterns that might emerge for other translation properties and an in-depth look at the interpretation of feature weights. Given the limitations of the study in terms of text and feature selection, repetition of this analysis on a different corpus such as the Dutch Parallel Corpus (Macken et al. 2011) would further support our exploratory findings.

We believe that the multivariate approach adopted here is not only very useful for understanding the nature of translations – because it supports the simultaneous investigation of a whole range of features that might affect the make-up of translations – but is also very promising for various other areas of the study of language variation.

Acknowledgments

The authors would like to thank Sascha Diwersy for extensive discussions of this research, which builds on joint work reported in Diwersy, Evert, and Neumann (2014). We would also like to thank the audience at the workshop "New ways of analyzing translational behavior in corpus-based translation studies" as well as the reviewers for valuable comments. Part of the research reported here was funded by the German Research Council under grants no. STE 840 / 5-2 and HAN 5457 / 1-2.

Appendix

The linguistic features in alphabetical order

(cf. Neumann (2013), see Hansen-Schirra, Neumann and Steiner (2012, chapter 3) for a full description of the annotation referred to here)

adja_T: attributive adjectives, per no. of tokens
English: all tokens receiving the part-of-speech tag "JJ.*" (general adjective), computed as the proportion of all tokens per text.
German: all tokens receiving the part-of-speech tag "ADJA" (attributive adjective), computed as the proportion of all tokens per text.

colloquialism_T: colloquialisms per no. of tokens
English: all strings like *yeah, bloody, damned, bitch, sissy, crap, buddy* etc., computed as the proportion of the total number of tokens per text
German: all strings like *toll, spitze, geil, bekloppt, bescheuert, Weichei, Blödmann, Klamotten* etc., computed as the proportion of the total number of tokens per text.

contractions_T: contractions per no. of tokens
English: all strings like *'m, 's, 't* etc. and, where applicable, a part-of-speech tag like "P.*" (pronoun) followed by a string like *'s, 'll* etc., computed as the proportion of the total number of tokens per text.
German: all strings like *gibts, willste, biste, guck, kuck, mal, drauf, runter, sagste, rüber, aufs, ums, nebens, 'n* etc., computed as the proportion of the total number of tokens.

coordination_T: coordinating conjunctions, per no. of tokens
English: all tokens receiving the part-of-speech tag "CC" (coordinating conjunction), computed as the proportion of the total number of tokens per text.
German: all tokens receiving the part-of-speech tag "KON" (coordinating conjunction), computed as the proportion of the total number of tokens per text.

finites_S: finite verbs, per no. of sentences
English & German: all items receiving the tag for finite verb (chunk_gf="fin") in the manual grammatical annotation, computed as the proportion of the total number of sentences per text.

imperatives_S: imperative mood, per no. of sentences
English: all sentences starting with the part-of-speech tag "VV0" (manually verified), computed as the proportion of all sentences per text
German: all sentences (manually verified) starting with the part-of-speech tag "VVIMP" (for the German imperative verb mood), and "VVFIN" ending on *-en* (for the plural form) followed by the personal pronoun *Sie* (for polite imperatives), and the part-of-speech tag "VV.*" ending on *-n* at the end of a sentence (represented by a punctuation mark) as the only verb in the sentence, computed as the proportion of all sentences per text.

interrogatives_S: interrogative mood, per no. of sentences
English & German: all sentences (manually verified) ending with a question mark, computed as the proportion of all sentences per text.

lexical.density: lexical density
English & German: all lemmatized items assigned a part-of-speech tag for nouns, full verbs, adjectives and adverbs computed as the proportion of the total number of tokens per text.

lexical.TTR: lexical type token ratio
English & German: all lemmatized items assigned a part-of-speech tag for nouns, full verbs, adjectives and adverbs computed as the proportion of the total number of items assigned a part-of-speech tag for noun, full verb, adjective and adverb tokens per text.

modal.adv_T: modal lexis per no. of tokens
English: all strings *very, highly, fully, completely, extremely, entirely, strongly, totally, perfectly, absolutely, greatly, altogether, thoroughly, enormously, intensely, utterly, only, almost, nearly, merely, hardly, slightly, partly, practically, somewhat, partially, scarcely, barely, mildly, just, really, most, more, quite, well, anyway, anyhow* in combination with the part-of-speech tag "R.*" (all adverbs), computed as the proportion of the total number of tokens per text.
German: all strings *sehr, ziemlich, recht, ungewöhnlich, höchst, außerordentlich, ziemlich, fast, nahezu, ganz, aber, vielleicht, denn, etwa, bloß, nur, mal, nun, nunmal, eben, ruhig, wohl, schon, ja, doch, eigentlich, auch, lediglich, allein, ausschließlich, einzig, ebenfalls, ebenso, gleichfalls, sogar, selbst, gerade, genau, ausgerechnet, insbesondere, erst, schon, noch* in combination with the part-of-speech tag "ADV" (adverb) and where applicable following the "V.*FIN" (finite verb), computed as the proportion of the total number of tokens per text.

modals_V: modal verbs per no. of verbs
English & German: all items receiving the part-of-speech tag "VM.*" (modal verb), computed as the proportion of the total number of verbs per text.

nn_T: nouns per no. of tokens
English & German: all items receiving the part-of-speech tag "N.*" (all nouns), computed as the proportion of all tokens per text.

nominal_T: nominalizations per no. of tokens

English: all tokens receiving the part-of-speech tag "N.*" and ending on *-ion, -ism, -ment, -ness* and the respective plural endings, computed as the proportion of all tokens per text.
German: all tokens receiving the part-of-speech tag "N.*" and ending on *-ung, -heit, -keit, ismus* and their respective plural endings, computed as the proportion of all tokens per text.

passive_V: passive voice, per no. of verbs
English: the results for the query for the part-of-speech tag "VB.*" followed by "VVN" (manually verified) with up to 3 intervening tokens, computed as the proportion of the total number of verbs per text.
German: the results for the query for strings of the auxiliary *werden* followed (or preceded) by "VVPP" (manually verified) with up to 8 intervening tokens, computed as the proportion of the total number of verbs per text.

past_F: past tense, per no. of finite verbs
English: all items receiving the tag for finite verb (chunk_gf="fin") in the manual grammatical annotation in combination with the part-of-speech tag "V.D.*" anywhere within the chunk, computed as the proportion of the total number of finites per text.
German: all items receiving the tag for finite verb (tns="past") in the morphology annotation, computed as the proportion of the total number of finites per text.

place.adv_T and time.adv_T: place and time adverbs, per no. of tokens
English: all tokens receiving the part-of-speech tag "RL" (adverb of place or direction) and "RT" (adverb of time), computed as the proportion of all tokens per text
German: all strings (and their variants in upper case) *hier, da dort, oben, unten, rechts, links, vorn.*, hinten, vor, dorthin, herab, herbei, dahin, jenseits, hinein, hierhin, hinunter, hierher; heute, jetzt, zuletzt, bald, sofort, morgen, derzeit, einst, früher, gestern, später, heutzutage, soeben, kürzlich, nachher, demnächst, jüngst, vorgestern, unlängst,* etc., computed as the proportion of all tokens per text.

prep_T: prepositions per no. of tokens
English: all items receiving the part-of-speech tag "I.*" (all prepositions), computed as the proportion of all tokens per text.
German: all items receiving the part-of-speech tag "AP.*" (all prepositions), computed as the proportion of all tokens per text.

pronouns_T: personal pronouns, per no. of tokens

English: all strings *I, me, mine, you, yours, he, him, his, she, her, hers, it, we, us, ours, they, them, theirs* in combination with the part-of-speech tag "PP.*" (all personal pronouns), computed as the proportion of all tokens per text.
German: all strings *ich, mir, mich, du, dir, dich, er, ihm, ihn, sie, ihr, ihn, es, wir, uns, euch, ihnen, Sie, Ihnen* in combination with the part-of-speech tag "PPER" (personal pronoun), computed as the proportion of all tokens per text.

subordination_T: subordinating conjunctions, per no. of tokens
English: all tokens receiving the part-of-speech tag "CS.*" (subordinating conjunction), computed as the proportion of the total number of tokens per text
German: all tokens receiving the part-of-speech tag "KOU.*" or "KO-KOM" (subordinating conjunction), computed as the proportion of the total number of tokens per text.

adv.theme_TH, obj.theme_TH, subj.theme_TH, text.theme_TH and verb.theme_TH: specific themes per no. of themes
English & German: the first grammatical function in each sentence (adv.theme: all adverbials, obj.theme: all objects and predicatives, subj.theme: subjects, text.theme: conjunctions and other types of connectives, verb.theme: verbs) in the manual grammatical annotation, computed as the proportion of the total number of themes per text.

titles_T: titles per no. of tokens
English: all strings like *Doctor, Professor, Sir, President, Senator, Chairman* etc., computed as the proportion of the total number of tokens per text.
German: all strings like *Doktor, Professor, Präsident, Minister, Botschafter* etc., computed as the proportion of the total number of tokens per text.

token_S: tokens per sentence

English & German: all token segments per text in proportion to all sentence segments per text.

References

Baker, M. 1993. Corpus Linguistics and Translation Studies. Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.), *Text and technology. In honour of John Sinclair*, 233–250. Amsterdam: John Benjamins.

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

Baroni, M. & S. 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.

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

Becher, V., J. House & S. Kranich. 2009. Convergence and divergence of communicative norms through language contact in translation. In K. Braunmüller & J. House (eds.), *Convergence and divergence in language contact situations*, 125–152. Amsterdam: John Benjamins.

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

Delaere, I. 2015. *Do translators walk the line? Visually exploring translated and non-translated texts in search of norm conformity*. Ghent: Ghent University PhD dissertation.

Delaere, I., G. De Sutter & K. Plevoets. 2012. Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target* 24(2). 203–224.

Diwersy, S., S. Evert & S. Neumann. 2014. A weakly supervised multivariate approach to the study of language variation." In B. Szmrecsanyi & B. Wälchli (eds.), *Aggregating dialectology, typology, and register analysis. Linguistic variation in text and speech*, 174–204. Berlin/New York: Mouton de Gruyter.

Hansen-Schirra, S., S. Neumann & E. Steiner. 2007. Cohesive explicitness and explicitation in an English-German translation corpus. *Languages in contrast* 7(2). 241–265.

Hansen-Schirra, S., S. Neumann & E. Steiner. 2012. *Cross-linguistic corpora for the study of translations - Insights from the language pair English-German*. Berlin: Mouton de Gruyter.

Hansen-Schirra, S. & E. Steiner. 2012. Towards a typology of translation properties. In S. Hansen-Schirra, S. Neumann & E. Steiner (eds.), *Cross-linguistic corpora for the study of translations - Insights from the language pair English-German*, 255–279. Berlin: Mouton de Gruyter.

House, J. 2002. Maintenance and convergence in translation – Some methods for corpus-based investigations. In H. Hasselgård, S. Johansson, B. Behrens & C. Fabricius-Hansen (eds.), *Information structure in a cross-linguistic perspective*, 199–212. Amsterdam: Rodopi.

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

Macken, L., O. De Clercq & H. Paulussen. 2011. Dutch Parallel Corpus: A balanced copyright-cleared parallel corpus. *Meta: Journal des traducteurs* 56(2). 374–390.

Mauranen, A. & P. Kujamäki, eds. 2004. *Translation Universals. Do They Exist?* Amsterdam: John Benjamins.

Mauranen, A. 2004. Corpora, universals and interference. In A. Mauranen & P. Kujamäki (eds.), *Translation universals. Do they exist?*, 65–82. Amsterdam: John Benjamins.

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

Oakes, M.P. & M. Ji (eds.) 2012. *Quantitative methods in corpus-based translation studies. A practical guide to descriptive translation research*. Amsterdam: John Benjamins.

Olohan, M. & M. Baker. 2000. Reporting *that* in translated English. Evidence for subconscious processes of explicitation?. *Across Languages and Cultures* 1(2). 141–158.

Serbina, T. 2013. Construction shifts in translations: A corpus-based study. *Constructions and Frames* 5(2). 168–91.

Steiner, E. 2008. Empirical studies of translations as a mode of language contact – 'Explicitness' of lexicogrammatical encoding as a relevant dimension. In P. Siemund & N. Kintana (eds.), *Language contact and contact languages*, 317–346. Amsterdam: John Benjamins.

Steiner, E. 2012a. A characterization of the resource based on shallow statistics. In S. Hansen-Schirra, S. Neumann & E. Steiner (eds.), *Cross-linguistic corpora for the study of translations - Insights from the language pair English-German*, 71–89. Berlin: Mouton de Gruyter.

Steiner, E. 2012b. Generating hypotheses and operationalizations. The example of *explicitness/explicitation*. In S. Hansen-Schirra, S. Neumann & E. Steiner (eds.), *Cross-linguistic corpora for the study of translations - Insights from the language pair English-German*, 55–70. Berlin: Mouton de Gruyter.

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

Toury, G. 2012. *Descriptive Translation Studies – and beyond: Revised edition*. 2nd ed. Amsterdam: John Benjamins.

Volansky, V., N. Ordan & S. Wintner. 2015. On the Features of Translationese. *Digital Scholarship in the Humanities* 30 (1): 98–118.

Notes

[1] The paper is also useful in providing a comprehensive overview of the state of the art of machine learning approaches to translationese.

[2] Implicitly, machine learning approaches – as well as our approach – adopt adherence to target language norms as the basis of comparison. However, from the point of view of translation studies it is not obvious to which norms translators should adhere, i.e. which translation strategy they adopt. Mimicking, as it were, original texts written in the target language is but one option, others being foreignization (e.g. induced by the perceived prestige of the source language), register norms (which are not the same as general source or target language norms), cultural expectations towards translations, specific translation briefs etc. For obvious reasons an individual, text-specific influence of the target language is impossible, but a more general influence of the target language could, for instance, mean that a translation in an aligned text pair displays a tendency to replace features of the source text untypical of the source language by features more typical of the target language. While this paper concentrates on the part of the variation in translation linked to shining through, our results also suggest a normalization effect in the translation direction German-English, which might be linked to target language influence (see Figure 3).

[3] In comparison to Diwersy, Evert and Neumann (2014) one additional feature was discarded because of collinearity. Another feature (the frequency of infinitives) had to be discarded because the automatically obtained frequency counts turned out to be unreliable.

[4] Note that only the first element in the sentence is analysed as the theme.

[5] As laid out in section 1 we include shining through as one of the properties of translation thus opposing to Baker's (1993) exclusion of source language interference.