A multidimensional approach to aligned sentences in translated text

Gard B. Jenset [1] and Lidun Hareide[2]

[1]University of Oxford; [2] University of Bergen

Abstract

Using unsupervised clustering techniques this study explores sentence alignment patterns in a parallel corpus of Norwegian source texts and Spanish translations, the NSPC (Hareide and Hofland 2012). The results show that three strategies with respect to sentence alignment dominate: one to one correspondence, merging two sentences into one, and removing sentences altogether (omission). The strategies are intricately correlated with the variables *translator*, *author*, and *genre*. However, we show how visualization techniques for cluster analyses offer a possibility for teasing apart these interactions as well as their relative importance. Our results indicate that non-fiction texts allow translators more freedom with respect to the treatment of sentences than do texts that are written by professional authors of fiction. The style of the author appears to play only a secondary role, but is especially important in fiction.

Keywords: corpus based translation, cluster analysis, parallel corpora, corpus alignment, unidirectional bilingual corpus

*** Principal contact:**
Gard B. Jenset, PhD
Visiting academic
University of Oxford, GB
Tel.: +447553789151
Email: gjenset@gmail.com

1. Introduction

Mainstream linguistics has been somewhat slow in appreciating and adopting a methodology based on corpus linguistics despite increasing focus on *usage-based* approaches (Gries and Divjak 2010). On the other hand, applied fields have caught on more quickly, and electronic corpora are now the mainstay of modern lexicography (Atkins and Rundell 2008), while in translation studies corpus-based methods are rapidly gaining ground (Oakes and Ji 2012). However, recent papers have criticized research in corpus based translation studies for not adhering rigorously enough to the methodology of corpus linguistics (De Sutter, Delaere, and Plevoets 2012:326), (De Sutter, Goethals, Leuschner, and Vandepitte 2012). They argue that researchers in Corpus Based Translation Studies tend to study only one variable, thereby demonstrating insensitivity to other factors such as text type and source language variation that may influence the variable studied or have explanatory power. The present paper presents a model of how a study analyzing multiple corpus variables can be conducted.

Traditionally the main purpose of an electronic corpus is to furnish the researcher(s) with distributional frequencies of words/lemmas and their contexts (Gries and Divjak 2010:338). However, in the present article we argue that the process of corpus creation itself (as opposed to merely the end product of it) can yield interesting and relevant data for translation studies. Our analysis of the results from creating an aligned unidirectional parallel corpus[2] *The Norwegian Spanish Parallel Corpus*, or NSPC, (Hareide and Hofland 2012), indicates that, given the appropriate quantitative techniques, corpora can provide researchers with broader and richer data than merely words and their contexts. Studies on alignment-patterns in parallel corpora have previously been conducted by Johansson (2011), Johansson & Hofland (2000) and by Ebeling & Ebeling (2013)(this volume).

In the context of translation studies, corpora offer an opportunity for teasing out the competing effects that authors, genres[3], translators and their language background have on the translated text. Such a complex problem requires an appropriate tool, and (Jenset and McGillivray 2012) argue that multivariate statistical techniques are particularly well suited for such studies (see for instance Gries and Wulff (Gries and Wulff 2012) and also other chapters in Oakes and Ji 2012 that adopt similar approaches). Although words or lemmas with their contexts are well suited for answering some questions, additional information can be gleaned from the process of aligning sentences from source texts with their translations. This is based on the assumption that when translators encounter a sentence in the source text, they have a number of options available to them, such as keeping the original sentence, merging it with a neighboring sentence, or splitting it into two or more sentences. Since the *unit of translation* is considered to be the entire text itself rather than the individual sentence (Zanettin 2000: 107; Laviosa 1997: 296; Baker 1995: 249), we consider this a reasonable assumption to make. Furthermore, our assumption is that this process is neither random nor mechanistic, but is potentially subject to source and target language influence, generic conventions, stylistic choices, as well as translators' styles, something we will return to in section 4 below. Our aim is

[2] We use the definitions of a bilingual parallel corpus provided by Altenberg and Granger "original texts in one language and their translation into one or several other languages" (Altenberg and Granger 2002:7–8) and by Aijmer " a collection of source texts and their translations, aligned at the sentence level" (Aijmer 2009).

[3] A short note on genre is due here, as defining genres is inherently problematic. In the NSPC, the texts were classified according to genre using data from an external database in order to avoid any subjective evaluation by the researchers (see Hareide and Hofland (2012: 84-87)).

hence to investigate any systematic patterns of association between alignment patterns and variables related to the author, the genre, the translator, the source language and dates of publication. We hypothesize that systematic translation strategies with respect to alignment will manifest themselves as detectable associations in the corpus data, and aim at studying such associations with a data-driven exploratory approach.

2. Methodology

The data for this study are composed of corpus alignment patterns. The unit of study is the individual text level, that is, for each text in the corpus the alignment information was coupled with information about the following variables: author (name), authors' gender, authors' choice of Norwegian written standard[4], the text's genre[5], date of publication of the original, translators' name, translators' gender, the translators' mother tongue, and the date of publication of the translation (Hareide and Hofland 2012: 84-89). Exploratory techniques, including visualization techniques, clustering, and regression modeling, were employed to establish which of these variables have explanatory power, leaving us with the variables author, genre and translator. The current study therefore proceeds to study these three variables in greater detail (see also section 3 below). Section 2.1 discusses the corpus data in more detail, while section 2.2 provides information about the statistical techniques used in the analyses.

2.1 Alignment data

Alignment in the NSPC was briefly explored in a previous publication, where a pronounced difference between texts with respect to the proportion of sentences that correspond one to one was observed (Hareide and Hofland 2012). In 16 of the 31 texts in the first version of the corpus, the percentages of sentences that correspond one to one are above 90. This result is concurrent with results from Johansson and Hofland (2000), who report a strong tendency towards one to one correspondence. However in the NSPC the percentage of sentences corresponding one to one range from 98.96 in the manual OL1 *Velkommen*! (Linnesund 2005)/*Bienvenidos: manual para nuevos habitantes de Noruega* (Torgersen 2008)[6] translated by Maria Luna de Torres, to only 75.08 percent in Åsne Seierstad's journalistic text *Bokhandleren i Kabul* (2002)/*El librero de Kabul* (2003b) (AS4) translated by Sara Høyerup and Marcelo Covián. Interestingly, the only other text translated by this duo, 101: *Hundre og én dag-en reportasjereise* (Seierstad 2003c)/101: *Ciento y un días* (Seierstad 2004) (AS3), has the second lowest score with only 76.71 percent of the sentences corresponding one to one.

In addition, the two texts translated by the team Høyerup and Covian also differ from the rest of the sample with regard to the relationship between the number of sentences in the Norwegian original and in the Spanish translation. AS3 and AS4 have the two lowest figures with regard to the relationship between the number of sentences in Spanish (s) and Norwegian (n), 0.83 and 0.84 respectively. These percentages may not seem remarkable; however, if we calculate the difference in sheer numbers, we can observe that the target text in AS3 is 1477 sentences shorter than the 8474 sentences of the original. The Spanish version of AS4 is 972 sentences shorter than the 6134 sentences of the Norwegian original. In comparison, the longest book in the corpus, LC1, has 27,717 sentences in its original version, and the translation is only

[4] Norwegian has two official written standards; the majority standard *Bokmål* and the minority standard *Nynorsk.*

[5] See Hareide and Hofland (2012: 84-87).

[6] This text has several authors, and the Norwegian library catalogue BIBSYS the Norwegian original is listed as Linnesund et. al. whereas the Spanish translation is listed as Torgersen et. al.

884 sentences shorter (Hareide and Hofland 2012:103–105). Thus we consider it of interest to explore whether such correspondence patterns can be put down to translators' style, or whether they seem to be indicative of complex interactions between more variables.

2.2 Multivariate data analysis

A dataset can be considered multivariate when several measurements or observations have been made on several units (Rencher 2002:1). In our case, the units are the translated texts and the measurements consist of proportions of various alignment patterns, such as one to one, one to zero (the original sentence has been incorporate into another one), and one to two (the original sentence has been split into two sentences), etc. Such multivariate data lend themselves to visualization techniques that attempt to reduce the systematic variation in the data into patterns that can be plotted and interpreted by the researchers. The present study takes an exploratory approach, where the aim is to cluster texts based on alignment patterns.

The dataset consists of 31 rows, one for each of the texts that constitute the first version of the NSPC (Hareide and Hofland 2012:84). The dataset has 17 columns, so that for each text the following variables were recorded: percentage of one to one alignment patterns (1:1), percentage of zero to one alignment patterns (0:1), percentage of one to zero alignment patterns (1:0), percentage of one to two alignment patterns (1:2), percentage of two to one alignment patterns (2:1), percentage of one to three alignment patterns (1:3), percentage of three to one alignment patterns (3:1), the author of the text, its genre, a code indicating the translator or translator team, a short hand code for the text, the number of words in the Norwegian source text (WN), the number of words in the Spanish translation (WS), the mean number of words per sentence in Norwegian (SLN), the mean number of words per sentence in the Spanish translation (SLS), the register of the text, and the language variety of the Norwegian source text (*Bokmål* or *Nynorsk*). Table 1 below shows an excerpt of the data used for the cluster analysis.

Table 1 *The dataset has 31 rows, each corresponding to a text, and 17 columns, each corresponding to different types of recorded information. The table shows selected columns for the first six texts, including percentages of three alignment patterns (1:1, 0:1, 1:0), the author, genre, translator team, text, the number of words in the Norwegian source text and the mean sentence length (in words) of the source text.*

1:1	0:1	1:0	Author	Genre	Translator	Text	WN	SLN
91.86	0.00	0.58	Audun Engelstad	epilogue	KBAL	AE1	3681	19.90
90.69	0.31	0.58	Åsne Seierstad	journalism	CF	AS1	102544	10.90
89.22	0.03	0.52	Åsne Seierstad	journalism	CF	AS2	90491	12.10
76.81	0.37	2.43	Åsne Seierstad	journalism	SHMC	AS3	91170	10.80
75.15	1.15	3.00	Åsne Seierstad	journalism	SHMC	AS4	76172	12.40
82.30	0.00	11.96	Brit Bildøen	epilogue	KBAL	BB1	3468	16.00

Clustering techniques can broadly be defined as unsupervised techniques that seek to partition the data into groups, or clusters, based on systematic patterns in the data (Baayen 2008:118), with the number of clusters being left unspecified (i.e. estimated from the data rather than specified by the researcher). A number of techniques exist, c.f. Baayen (2008:118-164), and the choice of specific technique depends both on the nature of the data and the aims of the study.

The present study employs two different techniques, namely agglomerative hierarchical clustering, and principal component analysis (PCA). As the present study aims to demonstrate, any one such technique may not be sufficient to analyze complex data on its own. Rather, an informed combination of different quantitative techniques may yield better results than using one in isolation. All analyses were carried out with the statistical package R (R Development Core Team 2011). The hierarchical clustering was carried out the R package *pvclust* (Suzuki and Shimodaira 2011).

Agglomerative hierarchical clustering begins with single data points and groups them into progressively larger groups, based on their dissimilarity. Hierarchical clustering is well suited for teasing out correlations in multivariate data (Baayen 2008:139). In translation studies, (Ke 2012) has used hierarchical clustering to group translated material into clusters that closely resemble acceptability judgments made by experts. Hierarchical clustering analyses are typically presented in a tree-like format known as a *dendrogram* where similar items are grouped close to each other. An important question pertaining to such clustering techniques is whether they are supported by the data or merely an artifact of the clustering algorithm (Everitt and Hothorn 2006:254). We used the *pvclust* package in R since it supports a bootstrap approach, where the data are shuffled around and clustered over and over again, in this case 1000 times. The ensuing dendrogram plots can then be controlled for consistency by examining how many times a particular cluster is formed. This prevents the clusters from being biased by the order in which the data points are combined, and allows us to single out the clusters that are most strongly supported by the data. In the dendrogram plots below, clusters that are found in at least 95% of the bootstrap cluster solutions are highlighted by a rectangle.

PCA, on the other hand, attempts to find the best overall structure in the data, while simultaneously achieving the best grouping of row and column variables, so that correlated categories will appear close each other in the ensuing *bi-plot*. The variation in the data is portioned into orthogonal principal components (PCs), and ordered in terms of their explanatory potential, so that PC 1 has the highest explanatory value, PC 2 the second highest, etc. The technique thus offers an ordered view to the importance of variables, since the correlations present in PC 1 (by convention the horizontal axis of the bi-plot) are more prominent in the data than those found in PC 2. Conversely, categories clustered around the center of the plot have little explanatory potential. As with hierarchical clustering, it is necessary to evaluate the quality of the solution. For a sound interpretation as much as possible of the variation in the data should be captured in the first two dimensions, as a percentage of the total explained variation. For a more comprehensive discussion of PCA in the context of translation studies see Jenset and McGillivray (2012).

3. Results and discussion

The interaction between variables such as genre, author, and translator can best be studied when there is room for comparison, that is, when authors have published within more than one genre, and been translated by more than one translator (team). That does not mean that other data are not of interest, but that they merely provide a background to the associations we intend to study. Three authors, Lars Saabye Christensen (LC), Jostein Gaarder (JG), and Åsne Seierstad (AS), appear with more than one text in the corpus and hence provide an interesting test case regarding how the variables author, translator, and genre interact. The two texts by Saabye Christensen are both novels, and are both translated by the team Kirsti Baggethun/Asunción Lorenzo (KBAL), who incidentally also have translated all three of Gaarder's contributions to the corpus (one novel, one collection of short stories and one book for children). Seierstad's four texts are all categorized as journalism, and as mentioned in section 2.1., two of the texts are translated by the team Sara Høyerup/Marcelo Covian (SHMC), whereas the remaining two are translated by Carmen Freixanet (CF). In the following sub-sections we present the results of an initial exploration (3.1), before moving on to a step-wise clustering exploration of the data (3.2). The final sub-section (3.3) employs PCA to make relations between the variables more explicit, and discusses the resulting patterns.

3.1 Exploring the effect of text length

To identify the variables that might have some explanatory power we carried out a number of exploratory steps. As an initial step we wanted to assess the influence of a variable like length, both in terms of the number of words per text and the mean sentence length. Unsurprisingly, we found that the mean sentence length of the original was significantly positively correlated with mean sentence length in the Spanish translation. A similar relationship was found between the number of words in the source text and in the translation. In both cases, we found that the translations tended to have higher values, although the mean estimated difference was only about 1 word extra (for both mean sentence length and number of words).

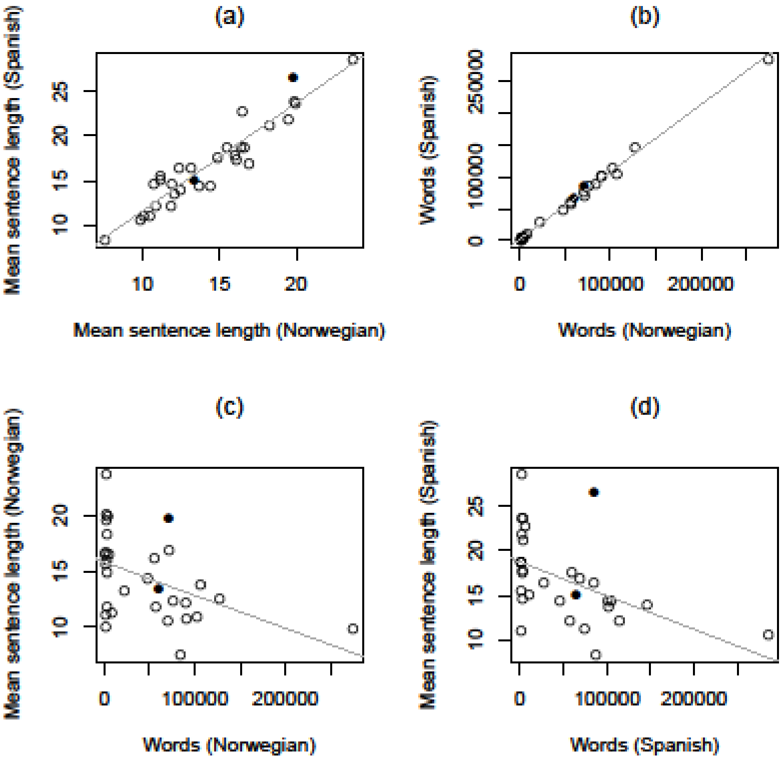

Figure 1 Scatter plots of length variables (WN, WS, SLN, SNS), with added regression lines. All regression analyses are statistically significant (p < 0.01). Circle types indicate source text variety (open: *Bokmål*, black: *Nynorsk*). No systematic differences in length appear to be correlated with variety. Panel 1a shows a very strong positive correlation between SLN and SNS (adjusted R^2 = 0.89). Panel 1b shows a similarly strong correlation between the WN and WS (adjusted R^2 = 0.99). Panel 1c shows that SLN and WN are negatively correlated. Although the result is significant the correlation is much weaker (adjusted R^2 = 0.19). A similar pattern can be seen in panel 1d for SNS and WS (adjusted R^2 = 0.21). Length properties are clearly preserved after translation.

We also found a significant but weaker correlation between the number of words in the source text and its mean sentence length, as well as a similar relationship between the number of words and the mean sentence length in the Spanish translations, cf. figure 1.

Since these variables were so closely connected, we carried out a hierarchical agglomerative cluster analysis based on these four variables (SLN, SLS, WN, WS), in order to identify any patterns among the texts. Figure 2 shows the result of this clustering. The bootstrap approach discussed in section 2.2 above allows us to confidently highlight the clusters that are best supported by the data (marked by a rectangle). As the left panel of the plot shows, the texts consistently fall into two groups. To get a clearer picture of what these groups represent, we re-created the plot, substituting the text codes with a binary variable, short vs. long. The variable indicates whether the Norwegian source text in question has a mean sentence length (SLN) above the mean (long), or below the mean (short). As the right hand panel shows, there is a perfect overlap between the two groups identified by the cluster analysis and the short vs. long labeling.

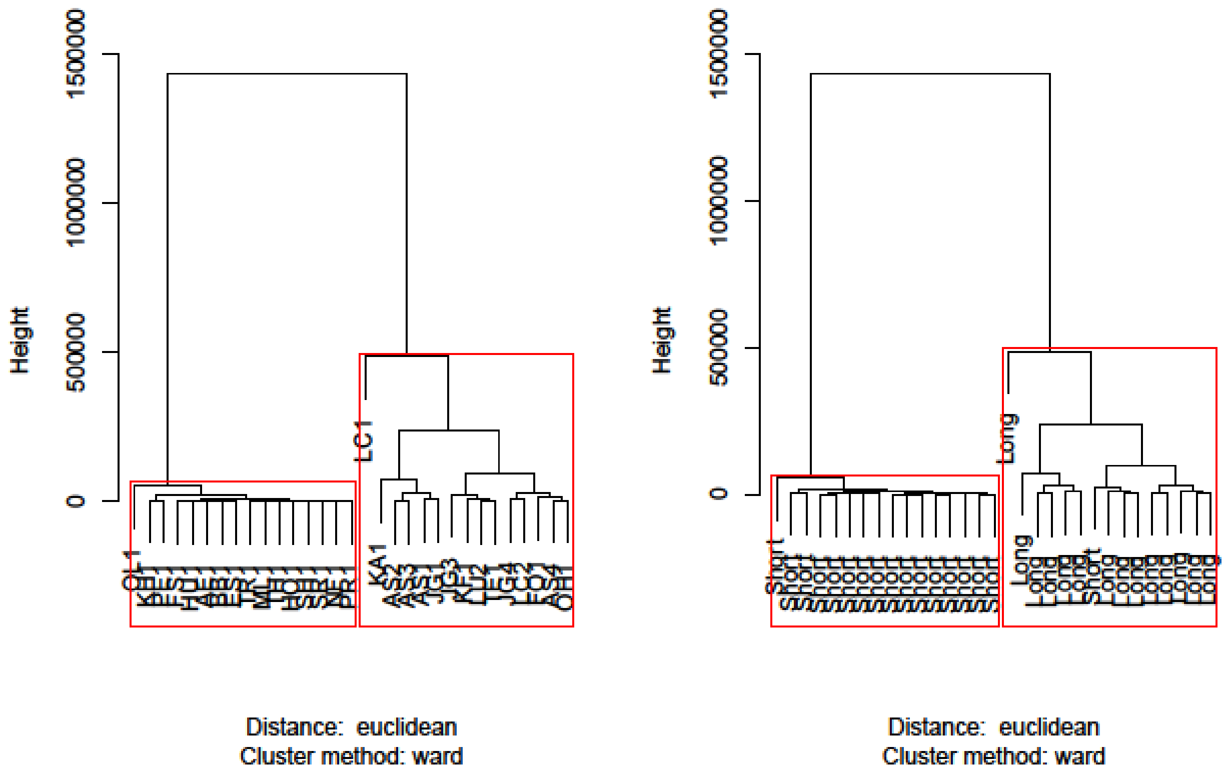

Figure 2 Hierarchical agglomerative cluster dendrogram of texts based on the four length variables (WN, WS, SLN, SNS). The left plot shows clearly that the texts can be reliably grouped into two clusters (highlighted by rectangles) based on length. The right hand plot illustrates this by plotting the same dendrogram, but with text labels replaced a binary variable indicating whether the Norwegian source text has a mean sentence length over the mean for the whole data set (*Long*) or below the mean (*Short*). Only a single text labeled as *Short* ends up in the right-hand cluster, indicating a very consistent correlation.

Based on the information in figures 1 and 2, we decided that it was reasonable to reduce the four length variables (WN, WS, SNL, SNS) to a single variable: short vs. long, based on the SLN variable. To identify any systematic relationships between length and the variables author, genre, and translator, we fit a binary logistic regression model, using the binary length variable as response and author, genre, and translator as predictors. None of the predictor variables were significant, which we take to indicate that no systematic differences in text length can be reliably associated with author, genre, or translator in our material.

3.2 Clustering rows with dendrograms

In the previous section, we established that text length was not related to the variables author, genre, and translator. Below, we will explore how these three variables are correlated based on the distribution of sentence alignment patterns. We scrutinize these variables one at the time using cluster dendrograms.

Figure 3 shows a hierarchical agglomerative cluster analysis of the sentence alignment data, with authors used as labels of the individual nodes (i.e. texts). As in the previous section, a bootstrap approach was used to identify the clusters that could be reliably supported by the data. Five clusters are highlighted, with four of them containing at least 3 texts, while the fifth only contains two of AS's texts. Figure 3 illustrates that the author variable alone is not sufficient to explain the variation in sentence alignment patterns. If the translators faithfully (or mechanically) reproduced the authors in this respect, we would expect clusters where works by the same authors were consistently clustered together. Instead, texts by AS appear in two different clusters, as do texts by Lars Saabye Christensen. Clearly, the author variable is an insufficient explanatory factor.

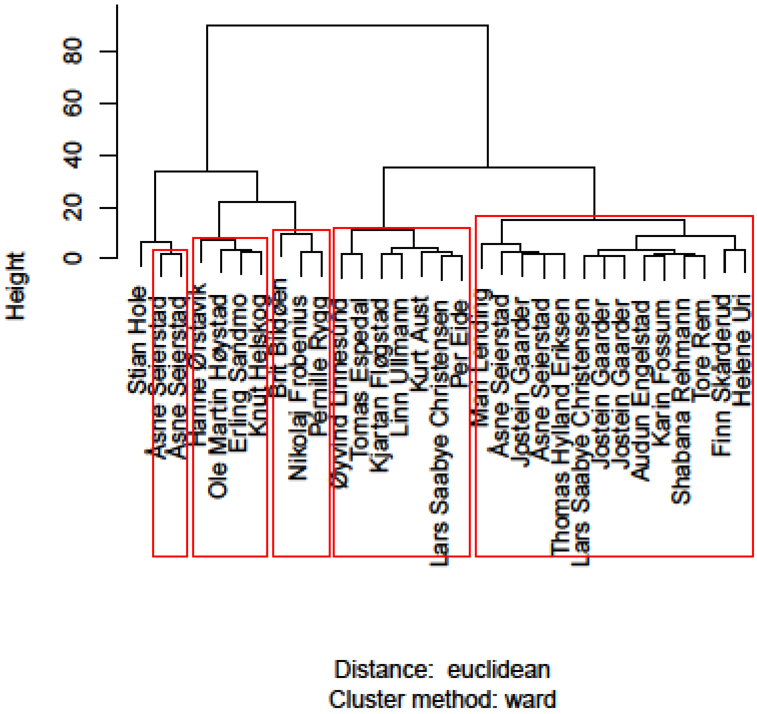

Figure 3 Hierarchical agglomerative cluster dendrogram of texts based on alignment patterns with leaf labels plotted as authors. The dendrogram is based on a bootstrap approach, and the best attested clusters are highlighted by a rectangle, indicating that the cluster appears in at least 95% of the bootstrap runs.

Plotting the exact same analysis but using genres as labels, as shown in figure 4, provides more clarity, as would be expected. For instance, the third cluster from the left consists only of epilogues, while the fourth cluster from the left is heavily dominated by novels. Nevertheless, considerable variation in genre remains within the clusters, and we must conclude that more variables need to be considered.

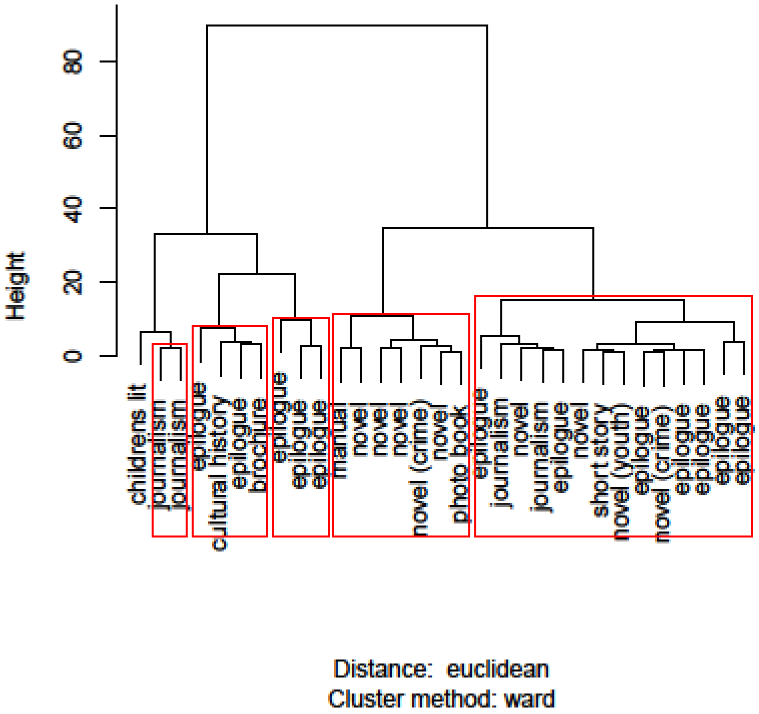

Figure 4 Hierarchical agglomerative cluster dendrogram of texts based on alignment patterns with leaf labels plotted as genres. Some degree of consistency can be observed within the clusters.

Turning instead to figure 5 we see the same analysis but this time plotted with translators as labels for the nodes. At first, figure 5 may seem to offer little more explanatory value than the preceding plots: translators are scattered across the clusters. However, a closer scrutiny reveals some systematic tendencies. The four texts by Åsne Seierstad, two in each main cluster, are now defined by their translators: two in the left-hand cluster translated by CF, and two in the right-hand cluster translated by SHMC. For the three texts by Jostein Gaarder, the picture is not as clear. All his texts have been translated by the same translator team (KBAL), a translator team that massively dominates the left-hand cluster. However, there is reason to doubt a one-variable explanation placing the explanatory burden on the translator variable. Most texts in the left-hand main cluster are translated by KBAL, yet some lower-level clusters can still be observed. More importantly, the right-hand cluster also contains texts translated by KBAL, illustrating that other variables must be involved. Consequently, we turn to a technique which allows us to study the interaction of these variables: PCA.

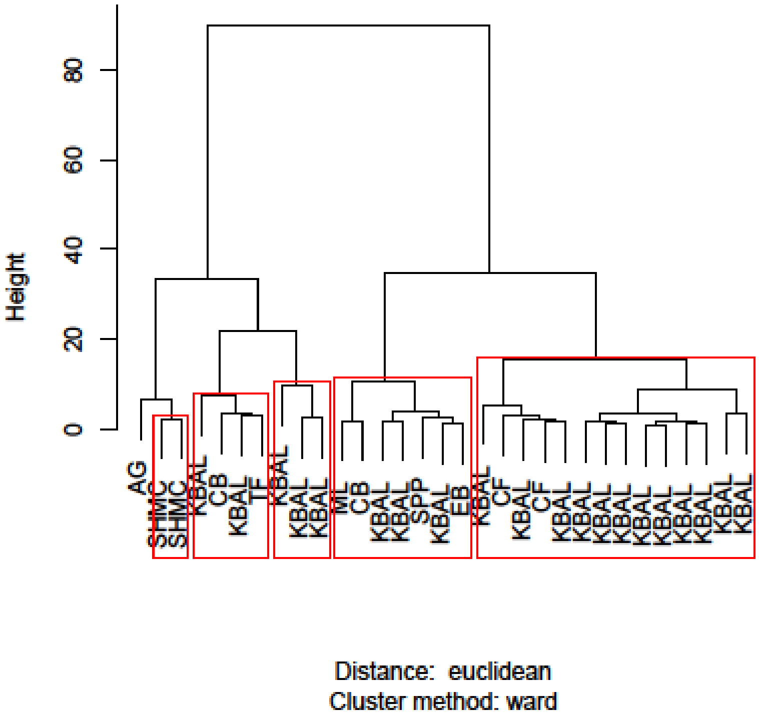

Figure 5 Hierarchical agglomerative cluster dendrogram of texts based on alignment patterns with leaf labels plotted as translators.

3.3 Exploring row-column interactions with PCA

As discussed in section 2.2 above, a PCA analysis allows us to capture more variables explicitly in the visualization than does a hierarchical cluster analysis. Although the underlying data are the same as for the hierarchical cluster analysis, the final visualization stresses the association between rows and columns, rather than focusing on rows alone. Figure 6 shows the resulting PCA bi-plot. The quality of the visualization is excellent, in that the first two dimensions (PC 1 and PC 2) together account for 94.5% of the variation in the data. The third-most important dimension, PC 3, accounted for only 4% of the variation, which was deemed to represent too little explanatory power to include in the final analysis (Baayen 2008:121).

Figure 6 illustrates that only three sentence alignment patterns (represented by the three labeled arrows which represent the dimensions in the plot) dominate the picture. The most important dimension in the figure is (by definition) the horizontal axis (PC 1). This dimension is dominated by the opposition between one to one patterns (1:1) and two to one (2:1) patterns, that is, between aligning translated sentences perfectly with their source or merging two Norwegian sentences to create one in Spanish. The fact that these two patterns are diametrically opposed to each other in the east – west direction, i.e. parallel to the horizontal axis, shows both that these two patterns are very different in their distribution but also that the difference between them is crucial to defining the horizontal (i.e. the most important) dimension. The vertical axis, which is (by definition) subordinate in importance to the horizontal axis, is dominated by the one to zero alignment pattern. This pattern clearly stands out with respect to the north – south axis in the plot, indicating that it differs from the one to one and two to one patterns.

An important property of PCA bi-plots is that we can visualize rows and columns in the same space. This allows us to interpret associations between row and column variables in terms of contiguity. If we were to draw a line from the center of the plot to each row point, we could calculate the angle between these lines and the lines representing the columns (i.e. the alignment patterns). If the angle is very small, the variables are strongly associated. An example of this is seen in figure 4 where 2:1 and SMHC are clearly strongly associated. For further discussion of the interpretation of bi-plots see e.g. Greenacre (2007:102–103) or Jenset and McGillivray (2012).

In figure 6 we focus on the variable that seemed most promising in the hierarchical cluster analysis, viz. translator. There seems to be some association between translators and alignment patterns. It is difficult to identify any clear correlations with respect to the one to one pattern; however, two to one patterns seem to be positively correlated with SHMG and AG. KBAL, the translator team responsible for the majority of texts in the corpus, predictably show up in the center of the plot when considering the horizontal axis. The second, vertical axis is dominated by the one to zero alignment pattern. Interestingly, while PC 1 showed variation among translators, PC 2 seem to document variation among the translations performed by the same team, viz. KBAL, since some of their translations make use of the one to zero strategy. However, since KBAL's translations are scattered over the whole vertical axis, there must clearly be more to the story.

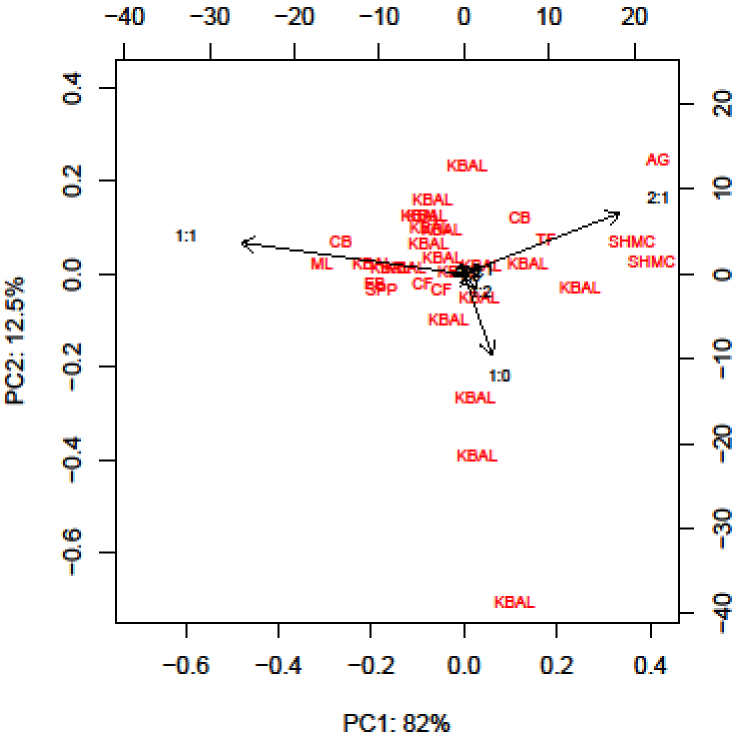

Figure 6 PCA bi-plot of the sentence alignment data. Three alignment patterns dominate. One to one and two to one (which are jointly responsible for PC 1) and one to zero, responsible for PC 2. Other alignment patterns have little or no impact on the analysis. The center of the plot is dominated by the most frequent translator-pair, but judging by PC 2, there is considerable variation in alignment even in the work of a single translator team. The visualization is excellent, with PC 1 accounting for 82% of the variation, PC 2 for 12.5%, bringing the total cumulative explained variance to 94.5%.

In figure 7, the same plot and data as in figure 6 is found, with one exception: the labels are now no longer translator names but genres. The overall interpretation of the plot is the same as for figure 6, but the plot in figure 7 offers a window to the effect of genres. Associated with the one to one alignment pattern we find genres like manuals, photo books, and some novels. The two to one alignment pattern, on the other hand, is associated with brochures, cultural history, journalism, and children's literature. Other types of writing, notably most novels, short stories, epilogues are placed in the center of the plot. Turning to the vertical axis, we find that the one to zero alignment patterns not only exhibits variation within a single translator team (as shown in figure 6), but also variation within a single genre, viz. epilogues. Again as with figure 6, we see that there is clearly unexplained variation in the plot.

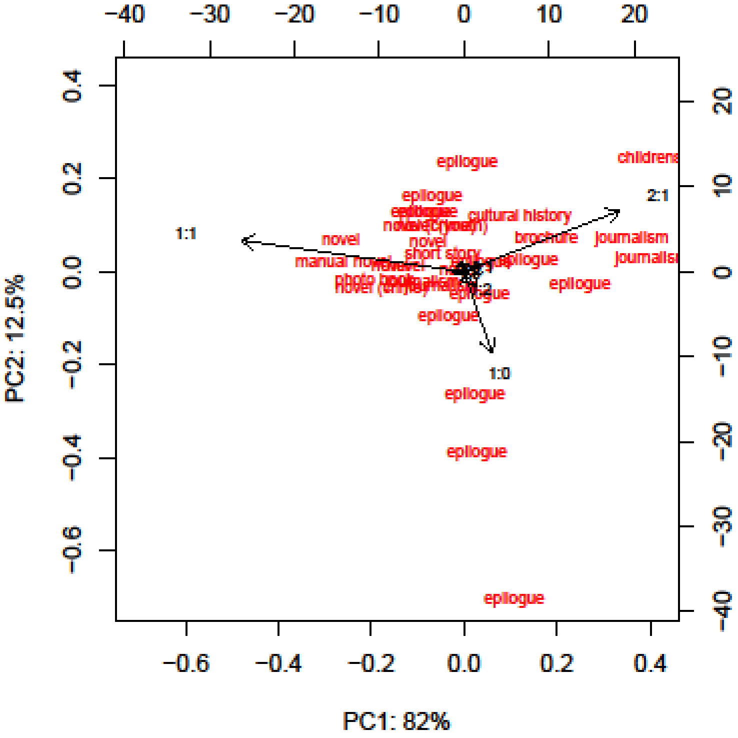

Figure 7 The same PCA bi-plot as in figure 6, but with genres as row labels. PC 1 appears to represent variation between genres, with a tendency for non-fiction to dominate the extremes of PC 1, whereas in PC 2 we can observe considerable variation within a single genre (epilogue).

Finally, in figure 8 we see a bi-plot with authors used as row labels. The unexplained variation observed along PC 2 in figures 6 and 7, where texts in the same genre and translated by the same team were scattered widely apart, seems to be a result of author variation. The 12 epilogues in the corpus all originate from the collection *Ytinger om Ibsen/Said about Ibsen* where twelve Norwegian authors and writers were invited to write epilogues to twelve of Henrik Ibsen's most important plays. The authors have very diverse backgrounds, from the Norwegian-Pakistani stand-up comedian and columnist Shabana Rehman, the professor of social anthropology Thomas Hylland-Eriksen to prize-winning authors of literature such as Brit Bildøen and Hanne Ørstavik, and the literature critics Mari Lending (who is also a professor of architecture) and Tore Rem (professor of English literature).

It is noteworthy that the authors who dominate PC 2 (Bildøen, Frobenius, and Rygg) all are professional authors of fiction. Conversely, other epilogues translated by the same team that were written by writers of non-fiction (e.g. Hylland-Eriksen, Rheman, and Skårderud) cluster around the middle. Thus, using PCA and plotting different type of metadata (translator, author, genre) we find that all the variables have varying degrees of explanatory power, and that these variables interact in subtle ways.

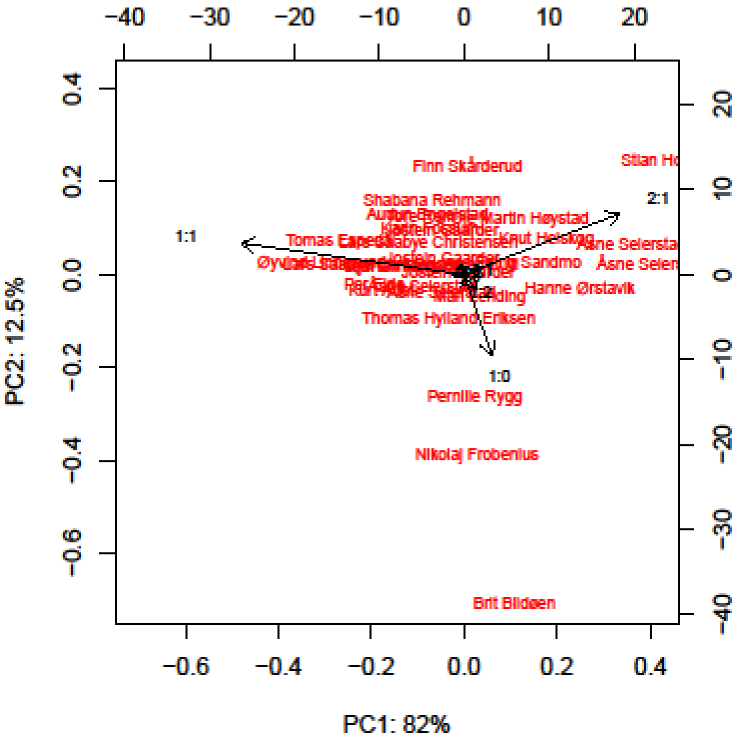

Figure 8 The same bi-plot as in figures 6 and 7, but with authors as row labels. The plot makes it clear that the residual unexplained variation along PC 2 observed in figures 6 and 7 can in fact be accounted for by differences between individual authors. Interestingly, the epilogue writers who dominate PC 2 all have a background as writers of fiction, whereas a number of prominent academics and columnists writing in the same genre gather around the center.

Based on the plots in figures 6 – 8 it seems safe to conclude that sentence alignment is correlated with all three initially proposed variables, viz. author, genre, and translator, but in subtly different ways. Focusing on the horizontal axis (PC 1), we see from figure 6 that it is correlated with differences between translators, but not differences within the works of translators (or translator teams) as attested by the concentration of KBAL in the center. Furthermore, from figure 7 we see a similar variation between genres, with non-fiction genres like manual, photo book, cultural history, brochure, and journalism making a heavy contribution. The vertical axis (PC 2) is associated with variation within translator teams -- notably KBAL -- (figure 6) and within genres -- notably epilogues -- (figure 7). Figure 8 reveals this variation to be largely due to differences between authors, in particular between epilogue writers with a background in fiction and those with a background in non-fiction.

We interpret this as follows: PC 1 primarily represents non-fiction and how those genres interact with translators. Put differently, the horizontal axis provides a picture of the maneuver space available to translators (or resulting from their informed judgment) when translating non-fiction prose with respect to strategies for dealing with sentences. PC 2 on the other hand represents fiction, and documents the role of the author and, possibly, the preservation of the authors' unique voice. As such, there appears to be more room or freedom for translators' judgment with respect to dealing with sentences when translating non-fiction than fiction. Since the principal components in a PCA are by definition independent from each other and ordered from the most important to the least important, we can draw some further conclusions. First, the interaction between genre and translator can be treated independently from the author variable. Second, the interaction between genre (non-fiction) and author (with respect to the final alignment patterns) of author and genre (fiction) can be described independently from the translator variable. Finally, the former interaction (genre and translator) is more important than the latter (genre and author).

4. Conclusions

The present paper has demonstrated that what may at first seem like a by-product of corpus-creation, the patterns of sentence alignment in a parallel corpus of translated text, can in fact provide insights about translated texts. As such, the study answers the call for more bottom-up methods in translation studies, taking more variables into account, essentially describing all the (relevant) parts of the corpus simultaneously. As the study has stressed, such an approach resting on many variables requires statistical techniques that are able to accommodate such data. As we have shown, such techniques are readily available in off-the-shelf statistical software and can easily be applied to material from translation corpora. One important benefit of our approach is that not only are we able to tease apart interactions of variables in the data, but we can also point the order, or order of magnitude, of such effects. Finally, such an approach is unbiased, in the sense that we can pick out what is more or less important with objective criteria arising from the data. In summary, the data we have studied indicate that both translator and genre affect sentence alignment, but that neither on its own has sufficient explanatory power. Taken at face value, this might seem like a trivial observation. However, the interpretation is based on an objective, bottom-up model, which means that the results arise from the data, rather than being imposed on them.

References

Altenberg, Bengt, and Sylviane Granger (eds.) 2002. Lexis in Contrast: Corpus-based approaches. Amsterdam: John Benjamins.

Aijmer, Karin. 2009. "Parallel and comparable corpora." In *Corpus Linguistics. An International handbook* edited by Anke Lüdeling and Merja Kytö, 275-292. Berlin/New York: Walter de Gruyter.

Atkins, Beryl T., and Michael Rundell. 2008. The Oxford guide to practical lexicography. Oxford: Oxford University Press.

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

Baker, Mona. 1995. "Corpora and translation studies: An overview and some suggestions for future research." Target no. 12:241-266.

Ebeling, Signe Oksefjell, and Jarle Ebeling. 2013. "From Babylon to Bergen: On the usefulness of aligned texts." In The Many Facets of Corpus Linguistics in Bergen, edited by Lidun Hareide, Christer Johansson and Michael Oakes. Bergen: Bergen Language and Linguistics Studies, BeLLS.

Everitt, Brian S, and Torsten Hothorn. 2006. A handbook of statistical analyses using R. Boca Raton, Fl.: Chapman & Hall/CRC.

Greenacre, Michael. 2007. Correspondence analysis in practice. 2nd ed. Boca Raton, FL.: Chapman & Hall/CRC.

Gries, Stefan Th, and Dagmar S Divjak. 2010. Quantitative approaches in usage-based cognitive semantics: myths, erroneous assumptions, and a proposal. Quantitative methods in cognitive semantics: corpus-driven approaches, ed by. by Dylan Glynn and Kerstin Fischer, 333–354. Berlin: Mouton de Gruyter.

Gries, Stefan.Th, and Stephanie Wulff. 2012. "Regression analysis intranslation studies." In Quantitative Methods in Corpus-Based Translation Studies, edited by Michael Oakes & Meng Ji. Amsterdam/Philadelphia: John Benjamins.

Hareide, Lidun, and Knut Hofland. 2012. Compiling a Norwegian-Spanish Parallel Corpus: methods and challenges. Quantitative Methods in Corpus-Based Translation Studies, ed by. by Michael P Oakes and Meng Ji, 75–113. Amsterdam: John Benjamins.

Jenset, Gard B, and Barbara McGillivray. 2012. Multivariate analyses of affix productivity in translated English. Quantitative Methods in Corpus-Based Translation Studies, ed by. by Michael P Oakes and Meng Ji, 301–323. Amsterdam: Jonn Benjamins Publishing Company.

Johansson, Stig. 2011. "Between Scylla and Charybdis. On individual variation in translation " Languages in Contrast no. 11 (1):303-328.

Johansson, Stig, and Knut Hofland. 2000. "The English-Norwegian Parallel Corpus: current work and new directions." In Multilingual corpora in teaching and research, edited by S.P. Botley, A.M McEnery and Andrew Wilson, 134-147. Amsterdam / Atlanta, GA: Rodopi.

Ke, Shih-Wen. 2012. Clustering a translational corpus. Quantitative Methods in Corpus-Based Translation Studies, ed by. by Michael P Oakes and Meng Ji, 149–174. Amsterdam: John Benjamins.

Laviosa, Sara. 1997. "How comparable can 'comparable corpora' be?" Target no. 9 (2):289-319.

Oakes, Michael P, and Meng Ji (eds.) 2012. Quantitative Methods in Corpus Based Translation Studies. Amsterdam: John Benjamins.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. Vienna. http://www.r-project.org.

Rencher, Alvin C. 2002. Methods of Multivariate Analysis. 2nd ed. New York: Wiley-Interscience.

De Sutter, Gert, Isabelle Delaere, and Koen Plevoets. 2012. Lexical lectometry in corpus-based translation studies: Combining profile-based correspondence analysis and logistic regression modeling. Quantitative Methods in Corpus-Based Translation Studies, ed by. by Michael P Oakes and Meng Ji, 325–345. Amsterdam: John Benjamins.

De Sutter, Gert, Patrick Goethals, Torsten Leuschner, and Sonia Vandepitte. 2012. Towards methodologically more rigorous corpus-based translation studies. Across Languages and Cultures 13.137–143.

Suzuki, Ryota, and Hidetoshi Shimodaira. 2011. pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. http://CRAN.R-project.org/package=pvclust.

Zanettin, Federico. 2000. "Issues in Corpus Design and Analysis." In Intercultural faultlines Research models in Translation Studies 1Textual and Cognitive Aspects, edited by Maeve Olohan. Manchester, UK &Northampton MA: St. Jerome Publishing.