From Babylon to Bergen: On the usefulness of aligned texts

Signe Oksefjell Ebeling and Jarle Ebeling

University of Oslo

Abstract

After outlining a short and select history of (the usefulness of) parallel texts and alignment, this paper presents a case study where the point of departure is a Norwegian text extract aligned against its translations into seven different target languages, using the Translation Corpus Aligner, originally developed by Knut Hofland. Our main concern is cases where there is not a one-to-one correspondence at sentence level between original (source) and translation (target) text. We seek to answer questions such as why a translator, translating into a specific language has chosen to split, or merge, a sentence in the source texts, while translators, translating into other languages have chosen not to do so. The study shows that a multitude of contributing factors seem to be involved , including author and translator style, target language constraints and preferences and perhaps even country- or language-specific translation guidelines.

Keywords: alignment; parallel texts; contrastive analysis; corpora; translation strategies

1. Introduction

In this article we wish to pay tribute to Knut Hofland and his contribution to the field of corpus linguistics by focusing on his collaboration with the late Stig Johansson, particularly in connection with the development of the English-Norwegian Parallel Corpus (ENPC) and the Translation Corpus Aligner (TCA).

In 1977, Stig Johansson went to Bergen to attend an introductory course in humanities computing; here he met the young computational engineer Knut Hofland. This first meeting led to a long-lasting friendship and professional collaboration that has benefited the (corpus) linguistic community at large.[1] Both were instrumental in the completion of the Lancaster-Oslo/Bergen Corpus, a fact the name of the corpus bears witness to; Oslo = Johansson and colleagues at the University of Oslo and Bergen = Hofland and colleagues at the University of Bergen. This early collaboration resulted in two major publications on word frequencies: Hofland & Johansson (1982) and Johansson & Hofland (1989).

A few years later, another collaborative project was initiated: The English-Norwegian Parallel Corpus project, where the principal members of the research team were Stig Johansson,

[1] The nature and extent of their collaboration is "documented" in correspondence between the two from as far back as 1977. Interestingly, while the linguist (Johansson) communicates by means of type-written letters and memos, the engineer (Hofland) communicates by means of hand-written letters and notes.

Knut Hofland and the present authors. The aim of the English-Norwegian Parallel Corpus (ENPC) project was to produce a corpus for use in contrastive analysis and translation studies. The model devised for such a corpus can be found in several publications,[2] but first appeared in Johansson & Hofland (1994) and is repeated here as figure 1.

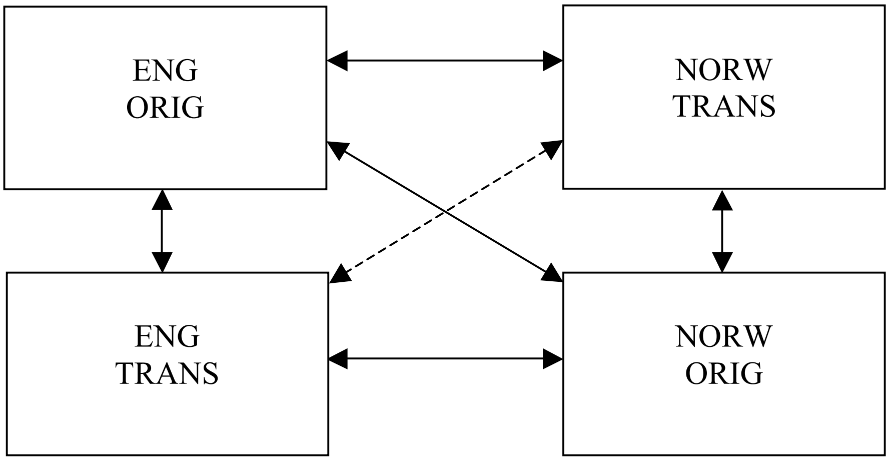


Figure 1. The structure of the ENPC (Johansson & Hofland 1994: 26)

The structure illustrates a corpus containing original texts and their translations (English to Norwegian and Norwegian to English). The fact that corresponding sections, e.g. orthographic sentences, of the source and target texts needed to be easily retrieved demanded an alignment program "*to specify equivalent points in the original and the translation*" (Johansson, Ebeling and Hofland 1996: 87). For this purpose, Knut Hofland developed the Translation Corpus Aligner (TCA).[3]

In the early 1990s, alignment of parallel texts was seen as a fairly new enterprise, although "*[by] the time the ENPC pilot project started in spring 1993, a good deal of work had already been done on alignment*" (Hofland 1996: 166).[4] In this article we will take the opportunity to paint a broader picture, hoping to illustrate that there is in fact a long-standing tradition of alignment and parallel texts spanning a period of more than 4000 years – from Babylon to Bergen, so to speak.

In addition to offering an overview of parallel texts and alignment, our aim is to use the TCA to complement previous studies carried out by Johansson & Hofland (2000) and Johansson (2011). The former produces some basic statistics of sentence divisions in original vs. translated texts from English into German and Norwegian, while the latter focuses on individual variation in translation, comparing alignment output of two short English texts and their ten different translations into Norwegian. The present paper takes a closer look at the aligned output of a Norwegian original text extract and its translations into seven different languages (Danish, English, French, German, Portuguese, Spanish and Swedish), enabling us to produce sentence division statistics reminiscent of Johansson & Hofland (2000) and to study translators' individual variation reminiscent of Johansson (2011). An additional question that is raised is whether it is possible to judge from the data at hand whether it is the target language that determines the choices rather than the individual translator?

[2] E.g. Johansson & Ebeling (1996), Johansson (1998), Johansson et al. (1999/2002), Johansson (2007).

[3] "The Translation Corpus Aligner was crucial in the building of the corpus" (Johansson, Ebeling and Oksefjell 1999/2002).

[4] Hofland (1996) mentions work by Brown et al. (1991), Gale & Church (1991) and Simard et al. (1992).

2. A condensed (and select) history of parallel texts

In the field of corpus-based contrastive linguistics a parallel text consists of two (or more) texts placed alongside each other in such a way that they can be easily compared. In our context the two texts will often be a source text (the original) and target texts (its translations). A compilation of such texts constitutes a parallel corpus. Most parallel corpora contain (pairs of) texts aligned at paragraph or sentence level. However, parallel texts may also be aligned at page or even text level.

Another kind of parallel text is one where the same event is described in several languages, but where there is no definite way of knowing whether one of the descriptions can be seen as the source of the other, parallel, descriptions (texts). The Rosetta Stone[5] is usually perceived as a parallel text of the first kind (cf. Véronis 2000: 1), but since there are some discrepancies between the three versions of the event described and the fact that the stone is severely damaged, we cannot be 100 % sure if it is not a parallel text of the second kind.

A prime example of a parallel text of the second kind is the Behistun Inscription, which is arguably as important as the Rosetta Stone when it comes to the deciphering of ancient languages and man's ability to read extinct scripts: "*The inscription includes three versions of the same text, written in three different cuneiform script languages: Old Persian, Elamite, and Babylonian (a later form of Akkadian). In effect, then, the inscription is to cuneiform what the Rosetta Stone is to Egyptian hieroglyphs: the document most crucial in the decipherment of a previously lost script.*" [6]

Both of the above texts can be seen as a form of royal propaganda, where the (new) ruler asserts his right to reign. In this respect, these ancient examples of parallel texts are different from e.g. the Hexapla,[7] where the texts have been collected and put together to perform text-critical studies of different editions of Bible texts, i.e. a parallel text of the first kind.

Note that the parallel texts just mentioned are different from texts making up comparable corpora, defined by McEnery & Hardie (2011: 20) as

> a corpus containing components that are collected using the same sampling method, e.g. the *same proportions* of texts of the *same genres* in the *same domains* in a range of *different languages* in the *same sampling period*.

The texts of the Rosetta Stone and the Behistun Inscription do not meet this definition, rather the parallelism lies in the fact that the same event is described in several languages. At a stretch, if we were to regard both the Rosetta Stone and the Behistun Inscription as being royal propaganda, they could, together with other royal propaganda texts of the same period, be part of a comparable corpus.

Despite the fact that the Rosetta Stone, the Behistun Inscription and the Hexapla contain the "same" text in different languages (and scripts), it is doubtful if this increased the number of people who could actually read these texts by very many. Making texts available to a broader audience is, however, one of the main reasons for making parallel text editions of classical works. The Loeb Classical Library,[8] for instance, "*presents important works of ancient Greek and*

[5] "Rosetta Stone." Encyclopædia Britannica. Encyclopædia Britannica Online. Encyclopædia Britannica Inc., 2012. Web. 09 Feb. 2012. <http://www.britannica.com/EBchecked/topic/509988/Rosetta-Stone>.

[6] Wikipedia contributors, "Behistun Inscription", Wikipedia, The Free Encyclopedia, 1 October 2012. <http://en.wikipedia.org/w/index.php?title=Behistun_Inscription&oldid=515401281>

[7] Wikipedia contributors. "Hexapla", Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 4 Oct. 2011. Web. 9 Feb. 2012. <http://en.wikipedia.org/w/index.php?title=Hexapla&oldid=453884474>

[8] http://www.hup.harvard.edu/collection.php?cpk=1031

Latin Literature in a way designed to make the text accessible to the broadest possible audience, by presenting the original Greek or Latin text on each left-hand leaf, and a fairly literal translation on the facing page."[9]

Coming back to our first kind of parallel text, where units, often orthographic sentences, referred to as s-units, are aligned, and where the purpose of aligning the two (or more) texts is to facilitate the study of the inter-relation between the texts or the languages of the texts, we should note that this also is a very old way of contrasting languages.

The text extract below is from a clay tablet from Assyria from the 2nd millennium BCE (see the Appendix for a drawing of the original tablet). It contains a parallel, aligned text of transliterated cuneiform signs, where the source text (in bold) is in Sumerian and the target text (in italics) is in Akkadian. (The […] and Xs represent damaged and unreadable segments and signs.)

1. […]	**X X […]**
2. […] X X X X	*par$_2$-ṣu-šu di X […]*
3. [lu$_2$] dam lu$_2$-da nu$_2$-a	**nam-tag-ga dugud-am$_3$**
4. ra-ḫu-u$_2$ aš-ti a-wi-lim	*a-ra-an-šu-kab-[tum-ma]*
5. lu$_2$ niĝ$_2$ nu-ĝar-ra gu$_4$ bal-e	**lu$_2$ eme-sig-ga g[u$_7$-gu$_7$-a]**
6. mu-ta-mu-u$_2$ nu-ul-la-a-ti	*a-kil kar-si*
7. lu$_2$ gaba-ri eĝer-ra-ni	**šu hul bi$_2$-in-du$_3$-a**
8. ša ar-ki mi-iḫ-ri-šu	*u$_2$-ba-an li-mut-ti i-tar-ra-ṣu[u$_2$]*
9. lu$_2$ dug$_4$-dug$_4$-ga nu-me-a im-ri-a-še$_3$	**mu-un-šub-ba**
10. ša$_2$ la qa$_2$-bi-tam el a-ḫi	*i-na-ad-du-u$_2$*
and so on …	

Figure 2. Text extract from an Assyrian clay tablet

An English translation of lines 3, 5, 7 and 9 would read (Lambert 1996):

3. Who has intercourse with (another) man's wife, his guilt is grievous
5. Who utters slander, who is guilty of backbiting
7. Who spreads vile rumours about his equal
9. Who lays malign charges against his brother

As can be seen from the layout of the transliterated text, and indeed the layout of the text on the original clay tablet, the alignment is of lines rather than s-units. However, since the text has a poetic flavour, it makes sense to align lines and not s-units or sentences.

It is difficult to know exactly why this tablet was produced, but we know from other sources that such bilingual tablets were produced by Akkadian-speaking students of Sumerian for learning purposes, either to learn to write Sumerian or to understand it. Sumerian is a

[9] Wikipedia contributors. "Loeb Classical Library." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 26 Jan. 2012. Web. 9 Feb. 2012. <http://en.wikipedia.org/w/index.php?title=Loeb_Classical_Library&oldid=473406226>

language isolate, while Akkadian is a Semitic language (see Ohgama & Robson 2010 and references therein).

We now jump from the bilingual clay tablets of the second millennium BCE to the second millennium CE and Brian Harris' (1988) ideas about bi-texts, hyper-bi-texts and interlinear bi-texts, which we shall see, show striking similarities. To Harris, a bi-text should be thought of not only as a complete source text (ST, original) and a target text (TT, its translation), but also as combinations of words and segments within the two texts. Harris writes (ibid.: 8):

> Yet translators do not translate whole texts at one fell swoop. They proceed a little at a time, and as they proceed each spurt, each segment forms a fragment of bi-text in their mind. Bi-text retains this structure when it is recorded on paper or in a computer: that is to say, not only is the whole text a bi-text but each segment combines ST and TT.

Having thus stored electronically combined bi-texts in the form of segments of source and target text, Harris imagines a database of hyper-bi-texts, i.e. a hypertext base or translation memory, where the translator can search his or her own previous work. Such a hyper-bi-text system should include a search engine "*programmed in such a way that when it finds an occurrence of the word [one is looking for], it retrieves and displays the whole translation unit in which it occurs*" (ibid.: 9).

Without specifying how, Harris foresees an even more sophisticated system whereby one can search for units similar to the one one is after, i.e. a system where what is combined is not translation units, but meaning units. Harris concludes by defining the concept of hyper-bi-text:

> it is bilingual hypertext stored in such a way that each retrievable segment consists of a segment in one language linked to a segment in the other language which has the same meaning (ibid.).

As to the display of bi-texts Harris advocates an interlinear translation display, where each line of the target text is interlaced between the corresponding line(s) of source text making up an interlinear bi-text. However, Harris (ibid.: 11) says *"[t]here remains the serious problem with interlinear bi-text, that of aligning ST and TT."*

In a series of experiments in the late 1980s and early 1990s it was shown how parallel texts could be automatically aligned at sentence level using sentence length in characters and/or lexical anchoring, i.e. finding words that correspond in the two texts to be aligned.[10] The method of using the number of character per sentence on text pairs already aligned at paragraph level (Gale & Church 1991, 1993), though very simple, was surprisingly effective. Gale & Church (1993: 89) write:

> The model was motivated by the observation that longer regions of text tend to have longer translations, and that shorter regions of text tend to have shorter translations. In particular, we found that the correlation between length of paragraph in characters and the length of its translation was extremely high (0.991). This high correlation suggests that length might be a strong clue for sentence alignment.

At the 14th ICAME conference in Zürich in 1993, Johansson & Hofland (1994) proposed a method for aligning English and Norwegian parallel texts which can be seen as an amalgamation of several of the methods tried out in earlier experiments. Johansson & Hofland, though

[10] See Véronis (2000) for an overview of the many methods explored and experiments performed at the time.

recognising the effectiveness of using sentence length in character, wanted to explore a more linguistic method of aligning a pair of texts at sentence level and proposed to include an anchor word list as a central component. The anchor word list contains words and expressions "*where the correspondence between the languages could be expected to be rather good*" (Johansson & Hofland 1994: 30), and the object of using anchor words "*was to calculate an* anchor score *which can be used in sentence alignment, expressing the number of shared anchor words*" (ibid.: 31).

The bilingual anchor word lists we use today contain function words, numerals, frequent and stable content words and names of days, months, countries and languages. In addition to sentence length in terms of characters and anchor words, the newest version of the alignment program also takes special characters, e.g. %, ?, !, matching proper names and cognate words in the two languages into consideration when aligning sentences.

In an investigation of the effect of the anchor word list carried out for the language pair English-Portuguese (Santos & Oksefjell 1999) it became clear that the anchor word list is essential to the success of the alignment. In fact, six out of the sixteen texts used in the investigation could not be aligned without the anchor word list, thus suggesting that a language-dependent aligner such as the TCA was an important step in the development of aligners for parallel texts.

In this short, and admittedly, select history of parallel texts we have seen how the idea of aligning texts of different languages (and scripts) is almost as old as the advent of writing itself, and that it has been used as a technique for conveying meaning simultaneously in different languages, for a range of reasons, including linguistic ones, for thousands of years.

In the following we explore how aligned texts can be used to highlight differences between languages and between source and target text.

3. Background

Our point of departure in this part of the paper will be two articles which both report on some basic statistics between original (source) and translated (target) texts aligned with the Translation Corpus Aligner. In Johansson & Hofland (2000) a small multilingual corpus of six English source texts and their translations into German and Norwegian are studied, while Johansson (2011) performs studies of a multiple translation corpus consisting of two short English texts and their ten different (commissioned) translations into Norwegian.

The articles lend evidence to what has been observed earlier, namely that

> automatic alignment may be significant in highlighting cases where a less direct translation has been chosen, thus pinpointing differences between languages and perhaps even providing a window into the mind of the translator. (Hofland 1996: 177)

Johansson & Hofland (2000) demonstrate that, although translations into both German and Norwegian show a strong tendency towards a one-to-one correspondence between orthographic sentences (s-units), in original and translated texts, there is some discrepancy between the German and Norwegian translations in terms of splitting and merging, i.e. whether one sentence in the source text is split into two or more sentences in the target text or vice versa. Similarly, Johansson's (2011) study shows that "*the number of sentences does not vary much across translations and is about equal in the translations and the original text*" (ibid.: 15). With regard to sentence division, Johansson finds that this is a fairly rare phenomenon, although "*sentence splits are consistently more common than merges*" (ibid.: 16). These observations point to some possible differences between languages and little individual variation between translators when it comes to sentence division. However, the fact that the number of sentences is "*slightly, but consistently, higher in the translations than in the corresponding original text [...] may reflect a tendency towards simplification in translated text*" (Johansson & Hofland 2000: 140).

Based on the corpus used for the present investigation we will compare the number of s-units in the original Norwegian text extract with translations into seven different languages, seeking to complement both Johansson & Hofland (2000) by looking at translations into more languages, i.e. seven, and Johansson (2011) by investigating how the same text is translated into different languages (by different translators).

Johansson (2011) finds that the Norwegian translations generally contain fewer word tokens than the original English texts, while the original texts contain a lower number of word types. We will also investigate the type-token ratio between original and translated texts, adding the dimension of a variety of languages in addition to a variety of translators. Moreover, as pointed out by Johansson (2011: 17) "*[e]arlier studies of translations of the same text have generally dealt with translations produced at different times*". In this study, not only were the translations produced shortly after the original was written, they are also translations of the same text into several languages.

Moreover, Johansson shows that, although the translators do not differ greatly in number of s-units produced, "*the extent of variation is very high*"; in fact, "*there are few sentences which are translated in an identical manner by different translators*" (ibid.: 4). In the present study we will not go into any great detail with regard to individual variation on the part of the translators, although we acknowledge that the question of variability among translators is an important one. This, in turn, should be seen in relation to the question of source-language vs. target-language orientation within translation studies, i.e. whether the translator belongs to a domestication or foreignization tradition. On the whole, however, we may expect more similarity between source and target texts from closely related languages.

This paper focuses on three main issues that can be directly or indirectly compared to results from the articles discussed above:

Type-token ratio between one Norwegian source text and translations into seven target languages

Number of s-units in one Norwegian source text and translations into seven target languages

Changes in sentence division between one Norwegian source text and translations into seven target languages

Will an investigation of translated texts of different languages show the same tendencies as those reported in Johansson & Hofland (2000) and Johansson (2011)?

Both the multilingual corpus used in Johansson & Hofland (2000) and the multiple translation corpus used in Johansson (2011) may be said to be closely related to the English-Norwegian Parallel Corpus project for which the alignment tool TCA was originally developed. In both articles it is the output from the TCA that is under scrutiny. As the alignment program has been described elsewhere, we will not go into detail as to how it works here. Instead we refer to Hofland (1996) and Hofland & Johansson (1998) for the original version of the TCA, and Izquierdo, Hofland and Reigem (2008) and Hareide & Hofland (2012) for its more recent, interactive version, TCA2, developed by Øystein Reigem.

4. Case study

For the case study we chose a Norwegian novel by Anne Holt that we knew had been translated into a number of languages. An extract of around 12,600 running words and its translations into

Danish, English, French, German, Portuguese,[11] Spanish and Swedish were prepared following the process applied in the ENPC project.[12] The TCA2 was used for the alignment of the parallel texts and an example of an s-unit in the original with its translations is given in example (1).

(1) Mellom linjene i et stort anlagt intervju med tre kjente psykologer og en pensjonert politimann fra Bergen, kunne man lese at morderen sannsynligvis var å finne blant utstemte Robinson-deltakere, mislykkede Idol-sangere eller tapende Grand Prix-finalister. (AnHo1N)[13]

Mellan raderna i en stort upplagd intervju med tre kända psykologer och en pensionerad polisman från Bergen kunde man läsa att mördaren troligen fanns bland utröstade Robinson-deltagare, misslyckade Idol-sångare eller förlorande Melodifestival-finalister. (AnHo1TSw)

Mellem linjerne i et stort anlagt interview med tre kendte psykologer og en pensioneret politimand fra Bergen, kunne man læse, at morderen sandsynligvis skulle findes blandt udstemte Robinson-deltagere, mislykkede Idol-sangere eller tabende Grand Prix-finalister. (AnHo1TD)

Einem großangelegten Interview mit drei bekannten Psychologen und einem pensionierten Kommissar aus Bergen konnte man entnehmen, daß der Mörder vermutlich unter ausgeschiedenen Robinson-Teilnehmern, mißglückten "Norwegen sucht den Superstar"-Sängern oder letztplazierten Grand Prix-Teilnehmern zu finden sei. (AnHo1TG)

The underlying implication of a major opinionated interview with three well-known psychologists and a retired policeman from Bergen was that the murderer was probably to be found among the ranks of people voted out of the Big Brother house, unsuccessful Pop Idol contestants or Eurovision Song Contest finalists who had n't won. (AnHo1TE)

Entre les lignes d' une vaste interview avec trois psychologues de renom et un policier berguenois en retraite, on lisait que le meurtrier pouvait probablement être retrouvé parmi les participants éconduits de "Robinson", les chanteurs manqués d'"Idol" ou les perdants à la finale de "Grand Prix". (AnHo1TF)

En una gran entrevista a tres psicólogos famosos y a un policía retirado de Bergen, se podía leer entre líneas que el asesino probablemente fuera uno de los concursantes descalificados de Robinson, de los cantantes fracasados del concurso Idol o de los finalistas perdedores de Eurovisión. (AnHo1TSp)

A insinuação subjacente a uma grande entrevista de opinião a três psicólogos de renome e a um polícia reformado de Bergen era a de que o assassino podia ser encontrado no seio dos grupos de pessoas expulsas da casa do Big Brother, de concorrentes malsucedidos do Ídolos ou de finalistas derrotados do Festival Eurovisão da Canção. (AnHo1TP)

[11] There are indications, such as using Johanne and Adam instead of Inger Johanne and Yngvar as names of the main protagonists, that the Portuguese translation is based on the English translation and not the original Norwegian text. Moreover, the title of the English translation is mentioned in the Portuguese version. Thus, the various numbers and statistics related to this text must be taken with a grain of salt.

[12] See e.g. Oksefjell (1999), Johansson (2007).

[13] AnHo1N stands for Anne Holt, text 1, Norwegian. Similarly, AnHo1TD stands for Anne Holt, text 1, translation, **D**anish and so on for **G**erman, **E**nglish, **F**rench, **P**ortuguese, **Sp**anish and **Sw**edish.

Following Johansson & Hofland (2000) and Johansson (2011) we will analyse material of the kind illustrated in example (1) in terms of some basic statistics mainly focusing on sentence division in original vs. translated text. However, the material used for the previous two studies and the present one are far from identical, and it is important to be aware of the differences and similarities between the three studies in terms of material, which can be summarised like this:

Johansson & Hofland (2000):
Number of source texts: 6
Source language: English
Target languages: German and Norwegian

Johansson (2011):
Number of source texts: 2
Source language: English
Target language: Norwegian
Multiple (10) translations of each of the two source texts

Current study:
Number of source texts: 1
Source language: Norwegian
Target languages: Swedish, Danish, German, English, French, Spanish and Portuguese
Multiple (7) translations of the same source texts

The differences are not without significance, and we will start by drawing attention to some relevant issues in this respect.

4.1 A note on language-specific issues

A major factor when counting tokens and types is how to treat contracted forms. Luckily there were no contracted forms in the Norwegian, Swedish, Danish and Spanish material and only one such form, *gibt's*, in the German version. As for English we followed the ENPC manual and split into two words forms such as *he's, she's, it's, he'll, he'd, won't, can't, let's* etc.[14] This was done automatically by the computer program developed for this purpose in the ENPC project.

The issue of where to split contracted forms gets more complicated when we turn to the French text, where both hyphens and apostrophes are used in writing to indicate contracted forms. Once more we turned to earlier work, work which was done in connection with the encoding of the French part of the Oslo Multilingual Corpus (OMC).[15] The compilation of the French texts for the OMC was a joint project with the University of Bergen, and Harald Ulland, associate professor of French language at that university, added rules of how contracted forms in French should be split. The main rule dictates that words should be split into two tokens whenever an apostrophe is encountered, e.g. *j'ai* → *j' ai* 'I have'. There are quite a few exceptions to this rule, e.g. *aujourd'hui* which should not be split into two. These exceptions where included in the program run on the French texts. In addition, pronouns are taken as separate word forms in cases such as *laissez-moi* → *laissez moi* 'let me', i.e. such cases were also split in the French text.

The issue of if and where to split contracted forms gets even more complicated when we turn to Portuguese, where the splitting of forms cannot always be done automatically where you find a hyphen. Particularly tricky are Portuguese verbs with enclitics and mesoclitics, i.e. atonic pronouns following the verb or occurring in the middle of the verb, respectively. Examples

[14] See http://www.hf.uio.no/ilos/english/services/omc/enpc/ENPCmanual.pdf
[15] http://www.hf.uio.no/ilos/english/services/omc/

include *dei-lho* 'gave-him-it' and fá-lo-ia 'do-it-would', i.e. 'would do it'. In order to ensure as comparable a result as possible with the other languages, we have chosen to regard the enclitics and mesoclitics as separate from the verbs they are attached to. This differs from the treatment proposed by Inácio & Santos (2008) in their guidelines for the annotation of the Portuguese part of the COMPARA corpus where they treat the verb + enclitic/ mesoclitic as one unit.

In addition to the issues discussed above, there are a host of morphological differences between the languages which we cannot address here and which make the token counts perhaps the least interesting, e.g. that definiteness is encoded by a suffix in the Scandinavian languages and that in Portuguese even names of persons may be preceded by the definite article.

4.2 Basic statistics

Once decisions regarding word splitting had been taken, relevant statistics were extracted for the data, including number of s-units, types, and tokens, in addition to information regarding splitting and merging of s-units in the source text as compared to the target texts. The alignment program shows its usefulness also with regard to this process. Not only does the TCA2 perform the actual alignment of the different text extracts, but the additional output it provides is instrumental in the present study. Figure 3 illustrates how one of the output files offers information regarding splits (1-2) and skips (1-0), for instance.

```
<link type='1-1' xtargets='AnHo1N.s50;AnHo1TE.s51'>

<link type='1-2' xtargets='AnHo1N.s51;AnHo1TE.s52 AnHo1TE.s53'>

<link type='1-1' xtargets='AnHo1N.s52;AnHo1TE.s54'>

<link type='1-1' xtargets='AnHo1N.s53;AnHo1TE.s55'>

<link type='1-0' xtargets='AnHo1N.s54;AnHo1TE.p20'>

<link type='1-1' xtargets='AnHo1N.s55;AnHo1TE.s56'>

<link type='1-2' xtargets='AnHo1N.s56;AnHo1TE.s57 AnHo1TE.s58'>

<link type='1-1' xtargets='AnHo1N.s57;AnHo1TE.s59'>
```

Figure 3. Extract from one of the TCA2 output files showing splits and skips

An overview of relevant statistics is given in Table 1, where the languages are listed according to geographical distance, i.e. Sweden is closer to Norway than Denmark, Denmark is closer than Germany, and so on.

Table 1
Basic statistics

	No.	**Sw.**	**Da.**	**Ge.**	**En.**	**Fr.**	**Sp.**	**Po.**
S-units	1,482	1,484	1,471	1,462	1,475	1,475	1,479	1,470
Tokens	12,599	12,679	13,109	13,421	14,123	15,380	14,344	13,591
Types	3,041	3,193	3,101	3,311	3,408	3,454	3,374	3,379
Split (1-2)	—	10	3	0	30	7	9	35
Merger (2-1)	—	5	7	17	31	15	21	38

Addition (0-1)	—	0	0	0	0	1	19	0
Skip (1-0)	—	3	5	2	7	0	4	8

Splits refer to cases where one s-unit in the source text has been split into two or more s-units in the target texts. By mergers are meant the opposite, i.e. where two or more s-units in the source correspond to one s-unit in the target. Additions refer to cases where an s-unit in the target texts has no correspondence in the source and skips are where one or more s-units in the source are left out in the target text.

With the exception of the Swedish translation, the translations have marginally fewer s-units than the Norwegian source text. This runs counter to the findings in Johansson & Hofland (2000), where the "*number of sentences was slightly, but consistently, higher in the translations than in the corresponding original text*" (Johansson & Hofland 2000: 140). A similar tendency was noted in Johansson (2011: 6) where, although "there is little variation in terms of the number of sentences", six of the translations have a slightly higher number than the original text, three have the same number of sentences, while one translation has a lower number of sentences. It is speculated that this slight increase in number of sentences in translation "*may reflect a tendency towards simplification in translated text*" (Johansson & Hofland 2000: 140). The fact that the opposite tendency in terms of number of sentences is noted here may suggest that the direction of translation may play a role. In both previous studies referred to the source text was English with translations into German and Norwegian (Johansson & Hofland 2000) and Norwegian (Johansson 2011). In other words, there may not be a general bias in translation for simplification in terms of number of sentences, as it may be dependent on the individual languages involved and the direction of translation.

With regard to number of tokens, Johansson (2011: 5) notes a "*considerable difference among translators*". It is tempting to draw a similar conclusion on the basis of the present material, as the difference between the translations into e.g. Swedish and French is of roughly 2,700 tokens. However, our material differs from Johansson's in that he investigates translations into the same language, while we are operating with translations into seven different languages. Thus, we are not only dealing with different translators, and a comparison is therefore seen as unfair due not least to the language differences commented on above, i.e. word splitting and morphological differences, and a greater tendency for compounding in Germanic languages in general. Interesting as this may be to follow up, it lies outside the scope of this article.

Although the number of types is perhaps not dependent on the nature of the individual languages to the same extent as tokens are, it is certainly related to the morphology of the languages under study here. Types reflect the variety of words used to cover the wording found in the source text. Interestingly, and apart from the Swedish and Danish material, there is a steady increase in the number of types used the further away (geographically) you get from Norway. This may have to do with the richer morphology of e.g. verbs in the Romance languages and also German and English when compared to Norwegian. Again, it seems to be pre-mature to suggest that individual translators differ in terms of types in our material, as the languages themselves dictate a greater variety of words to begin with. This calls for a more detailed study of how difference in morphology may influence the success of automatic alignment.

As would be expected, and as shown by Johansson (2011: 6-7) as well, most translators translate sentence by sentence, resulting in mainly 1-1 correspondences. This seems to hold also when translating from Norwegian into the Germanic and Romance languages under study here. However, as can be gleaned from Table 1, there are some conspicuously high numbers of 1-2 splits and 2-1 mergers for some of the languages. We will return to these below (sections 4.2.3 and 4.2.4). There are no instances of splits larger than 1-2 and very few mergers larger than 2-1, only six 3-1 and 4-1 mergers in total for all the languages, and even fewer 2-2 correspondences (three in total).

Although neither Johansson & Hofland (2000) nor Johansson (2011) address the issue of additions and skips in the translations, we have chosen to do so, since in Table 1 additions are shown to be a popular feature in the Spanish translations, while skips, although not very common, are found to be more evenly distributed across the translations. It will thus be interesting to see what seems to trigger the one or the other, and in the case of skips whether it is the same s-units that are omitted across the languages.

4.2.1 Additions

In the case of additions there is one target text that stands out, namely the Spanish one with 19 added s-units. The only other target text in which an s-unit has been added is the French translation. In example (2), the French translator has made explicit in a separate sentence that the delegate who was neither smiling nor laughing is a woman, '*Une femme*'. In the source text this is apparent from the use of the female possessive pronoun *hennes* 'her'. It could thus be claimed that this extra s-unit is a result of language-specific differences rather than a case of explicitation.[16]

(1) Blant alle de hujende, klappende og plystrende delegatene var det én som verken smilte eller lo. Hendene **hennes** beveget seg langsomt mot hverandre, i en demonstrativ, lydløs protest. (AnHo1N)

Parmi tous les délégués criant, applaudissant et sifflant, quelqu' un ne souriait pas, ne riait pas. **Une femme**. Ses mains allaient lentement l' une vers l' autre, en une protestation silencieuse, manifeste. (AnHo1TF)

In the Spanish translation, on the other hand, the additions seem to be instances of explicitation, in that the translator has chosen to add information that is not explicitly there in the source text. Examples include (3) and (4).

(2) — Finnes de egentlig, sa Yngvar oppgitt. — Profesjonelle mordere? Jeg mener, her i landet, i vår del av Europa?

Hun skakket på hodet og sendte ham et blikk som om han hadde spurt om det noen gang var vinter i Norge.

— OK, mumlet han. (AnHo1N)

— ¿De verdad hay de eso? — dijo Yngvar, hastiado de Asesinos Profesionales S.A. — Quiero decir, ¿en este país, en esta parte de Europa?

Ella ladeó la cabeza y lo miró como si hubiera preguntado si alguna vez era invierno en Noruega.

— ¿Lo preguntas en serio?

— Vale — murmuró él —. (AnHo1TSp)

(3) Et øyeblikk ble hun stående bak ham, tankefull, uten å fokusere på noe.

— Det er faktisk ikke nødvendig med flaks. (AnHo1N)

Se quedó un rato de pie detrás de él, pensativa, como sin enfocar en nada. **Dijo:**

— La verdad es que no le hace falta suerte. (AnHo1TSp)

[16] In the following we will use terms such as explicitation, simplification and normalisation, known as translation universals, taken from the field of translation studies (see e.g. Mauranen & Kujamäki 2004). Be aware, however, that our study is limited to looking at explicitation, simplification, etc. at the inter-sentential level and not at the intra-sentential level.

In example (3), the translator makes explicit the sarcasm implicitly present in the previous sentence in the source text, *som om han hadde spurt om the noen gang var vinter i Norge* 'as if he had asked if there was winter in Norway' by adding *¿Lo preguntas en serio?* 'Are you asking this seriously'. Some of the additions in the Spanish text are of this kind, while others are of the kind found in (4), where the translator has explicitly inserted a signal introducing direct speech: *Dijo* '(she) said'.

The fact that this kind of explicitation, by adding s-units, is only found in the Spanish text raises the question of whether this is a trait of translation into Spanish in general, or whether this is due to the preferences of one individual translator. To examine this in more detail, a corpus similar to the one in Johansson (2011) would be needed, including a number of Spanish translations of the same text.

4.2.2 Skips

Only a small number of skips was recorded overall in the material, ranging from zero in the French translation to eight in the Portuguese. Although the various translators do not leave out the same s-units, there is a tendency to leave out short, one-word s-units such as *nei* 'no' and *hvem* 'who', at least in the Swedish, Danish and German target texts. In the English and Portuguese texts in particular, and to some extent in the Spanish one, the skips include descriptive elements not part of the dialogue, e.g. example (5) where the two s-units in bold have been left out in the English and Portuguese target texts and example (6) where one s-unit has been left out in the Spanish target text.

(4) — Den ringte ikke, sa Inger Johanne fort. — Den pep. Og så ...

Yngvar fomlet med mobilen. Det grønne lyset blafret.[17]

— Herregud, mumlet han. (AnHo1N)

“It did n't ring,” Johanne explained quickly. “Just peeped. And then...”

“Jesus,” he mumbled. (AnHo1TE)

— Não tocou — explicou Johanne rapidamente. — Apenas apitou. Depois...

— Credo — balbuciou ele. (AnHo1TP)

(5) Motvillig rakte hun ham speilet. **Uttrykket hans vekslet fra vantro til fortvilelse.**[18]

— Jeg ser ut som et brød, jamret han. (AnHo1N)

Ella le pasó el espejo con gesto reluctante.

— Parezco un pan — se quejó Yngvar —. (AnHo1TSp)

Disregarding the Portuguese text, which we suspect is based on the English version, and focusing on what happens in the English and Spanish translations, we can perhaps conclude that the omission of descriptive units is a characteristic of two individual translators, and of one in particular as it happens more often in the English than in the Spanish target text.

4.2.3 Splits

The number of splits, i.e. where one s-unit in the source text is rendered by two s-units in the translations, ranges from zero for German to 38 for Portuguese. The tendencies with regard to what triggers the splits are summarised in Table 2.

Table 2

Splits: source text to target text processes

[17] Lit.: Yngvar fumbled with the mobile. The green light flickered.

[18] Lit.: His expression changed from disbelief to desperation.

ST → TT process Target language	semicolon in Norwegian → sentence division in TT	comma in Norwegian → sentence division in TT	... in Norwegian s-unit → sentence division with ... in TT[19]	other	Total
Swedish	8	—	—	2	10
Danish	—	3	—	—	3
German	—	—	—	—	0
English	9	15		6	30
French	4		2	1	7
Spanish	2	3	2	2	9
Portuguese	8	22	2	3	35

There are two major tendencies that can be noted, both of which involve the author's use of punctuation. The use of semicolon and comma, often with main clauses either side of the punctuation mark, seems to be the main reasons for the splits. It is tempting to conclude that it is a general trait of translated text to avoid semicolon and comma to combine main clauses. However, as will be seen in the section on mergers, commas at least are a popular device in the merging of s-units in translations. This leaves us with a notion that only the use of semicolons is under attack, as shown in example (6), where four of the seven translators have chosen a split (indicated by </s> <s> in place of the semicolon).

(6) <s>Tårene var store som **vanndråper; de** dvelte et sekund eller to i øyekroken før de løsnet og pilte ned i håret like under tinningen.</s> (AnHo1N)

<s>Tårarna var stora som **vattendroppar.</s> <s>De** dröjde en sekund eller två i ögonvrån innan de lossnade och tillrade ner i håret alldeles under tinningen.</s> (AnHo1TSw)

<s>Tårerne var store som **vanddråber; de** blev hængende et sekund eller to i øjenkrogen, inden de løb ud og randt ned i håret under tindingen.</s> (AnHo1TD)

<s> Die Tränen waren groß wie **Wassertropfen; sie** blieben ein oder zwei Sekunden im Augenwinkel hängen, dann lösten sie sich und verschwanden gleich unterhalb der Schläfe in den Haaren.</s> (AnHo1TG)

<s>Her tears were as big as **raindrops.</s> <s>They** gathered in the corners of her eyes for a moment before overflowing and running down her temples into her hair.</s> (AnHo1TE)

<s>Les larmes étaient grosses comme des **gouttes d' eau.</s> <s>Elles** s' arrêtaient une seconde ou deux au coin de l' oeil avant de se détacher et filer vers les cheveux juste sous la tempe.</s> (AnHo1TF)

<s>Las lágrimas eran grandes como **gotas de agua, y** se rezagaban un segundo o dos en el rabillo del ojo, antes de desprenderse y caer sobre el pelo de las sienes.</s> (AnHo1TSp)

[19] Three dots seem to be used by the author to indicate hesitation or an unexpressed thought on the part of the speaker.

> <s>As lágrimas eram do tamanho de **gotas de chuva.</s> <s>**Acumulavam- se nos cantos dos seus olhos antes de transbordarem e lhe escorrerem pelas têmporas abaixo, até ao cabelo.</s> (AnHo1TP)

The reason for this apparent dislike of semicolons among some of the translators should perhaps not be attributed to translation strategies such as simplification or explicitation. Rather it seems to be a case of overuse of this punctuation mark on the side of the author. The translators may have tried to keep their use to a less conspicuous level. If this is the case, normalisation may rather be the strategy at play, something that is reminiscent of Malmkjær's study of punctuation in H.C. Andersen's stories. According to Laviosa (2002: 56), Malmkjær "*provides evidence for shifts towards normalisation*" in that "*semi-colons become full-stops in the translations*". Indeed, "*[t]he author's [Andersen's] unusual style by Danish standards is therefore rendered more readable by clearly marking breaks in the information flow*".

4.2.4 Mergers

As mentioned in the previous section, commas in source and target texts seem to cancel each other out so to speak when splits and mergers are counted. Although a number of commas become full stops in the translations, the reverse is also true, where two s-units divided by a full stop in the Norwegian source text become one s-unit consisting of two clauses separated by a comma in the translations. These and other tendencies in terms of mergers in the translations are shown in Table 3.

Table 3
Mergers: source text to target text processes

ST → TT process Target language	comma added in TT	main clause + subclause/phrase in separate s-units in Norwegian → one sentence	... at the end of s-unit in Norwegian → one sentence with ...	coordination in TT	dash in connection with direct speech added	other	Total
Swedish	2	1	—	—	—	2	5
Danish	3	2	1	—	—	1	7
German	3	—	13	—	—	1	17
English	14	—	5	9	—	2	30
French	8	—	4	—	—	3	15
Spanish	11	—	5	—	—	5	21
Portuguese	20	—	2	5	5	6	38

The translation data for mergers is less clear-cut than was the case with splits. A more varied picture emerges, although the inclusion of a comma to join two clauses originally written as two full sentences separated by a full stop is a general tendency for all translators. Another trend with regard to mergers involves the author's use of three dots (...) at the end of an s-unit. After the three dots, the source text will often start with a full sentence, while the translation will have a lower case letter indicating that what follows is a continuation of what came before. This is particularly noticeable in the German text, where 13 of the 17 mergers are of this kind. The phenomenon is illustrated in example (8). Whether this is a sign of normalisation on the part of the German translator, in that German typically favours longer sentences than Norwegian, or whether it is this particular translator's preference, is hard to tell based on one target text only.

(7) <s>Eller **...</s> <s>Jo da**, hun kom litt tidligere.</s> (AnHo1N)

<s>Oder**... doch**, sie kam ein wenig früher zurück.</s> (AnHo1TG)

Another minor tendency that seems to be even more language (or translator) specific is mergers of two Norwegian s-units where only one is a main clause. In a few cases the Swedish and Danish translators opt for one s-unit consisting of a full sentence, perhaps in an attempt at normalising the syntax, e.g. example (9).

(8) <s>Hun gjorde ikke **det.</s> <s>Før** i tiden.</s> (AnHo1N) [20]

<s>Hon gjorde inte **det förr** i tiden.</s> (AnHo1TSw)[21]

Coordination within a sentence, or s-unit, seems to be a preferred choice particularly of the English translator. This is also noted for the Portuguese translator, albeit not to the same extent. In this context, it should again be stressed that the Portuguese translation may be a translation of the English version rather than the Norwegian one. Nevertheless, an example of this phenomenon in the English (and Portuguese) translation is shown in (10), and could perhaps be labelled explicitation, as it explicitly sees the relationship between the two Norwegian s-units as one of coordination.

(9) <s>Døren smalt igjen.**</s> <s>**Ragnhild skar i å gråte.</s> (AnHo1N)

<s>The door slammed **and** Ragnhild started to cry.</s> (AnHo1TE)

<s>A porta bateu **e** Ragnhild começou a chorar.</s> (AnHo1TP)

5. Conclusion

There are relatively few clear and consistent patterns that emerge from this small-scale investigation of sentence division in translated text across several languages. One reason for this may be the many variables present in the material, including:

individual author style
individual translator style
individual target language constraints
individual target language preferences
country-/ language-specific guidelines for translation

Nevertheless, it is tempting to conclude that there seems to be a geographical cline in connection with mergers, where Swedish and Danish, which are closely related languages, contain few mergers, while a steady increase is found the further away from Norwegian one gets, with some exceptions. It should again be noted that the English translation seems to be more influenced by the individual style of the translator, as was also the case with the skips; while they may be seen as indications of simplification in the target texts, it is only in the English (and Portuguese) and Spanish text that they are of a nature where textual content is lost. Similarly, perhaps the clearest tendency noted in the material may be attributed to the style of an individual translator is the Spanish text, where 19 additions, or inter-sentential explicitations, were found. In this respect, we note that among the ten translators discussed in Johansson's (2011), there was one or two who stood out as being different from the others in terms of basic statistics of the kind explored here.

[20] Lit.: She didn't do that. Earlier.

[21] Lit.: She didn't do that earlier in the time.

What may seem to be instances of explicitation in many of the translations due to splits are in a way cancelled out by mergers of the reverse kind, i.e. a split is triggered by the use of a comma in the source text, while a merger including a comma is used in the target texts.

To return to the question posed in the introduction of whether it is possible to judge from the data at hand whether it is the target language that determines the choices rather than the individual translator, our answer is both yes and no. Yes, in that we, in a couple of cases, have pointed to what may appear to be evidence of an individual translator's choice. No, in the sense that we have not been able to determine whether some of the sentence divisions are guided by language-specific constraints or preferences or by the individual translator. A case in point is the Swedish target text's bias towards sentence splitting rather than using the semicolon of the source text.

The fact that variables such as the ones pointed out above are part and parcel of translated material makes it clear that large translation corpora of different source languages and target languages, both of the kind investigated here and in Johansson & Hofland (2000) and in Johansson (2011), are needed. Only then will we be able to see more general tendencies for the languages involved. In this context it is also important to stress the usefulness of the alignment of parallel texts.

In fact, and as we have tried to illustrate in this article, alignment has proved its worth not only over the past few decades but over the millennia, as the idea of aligning texts stretches almost as far back as the advent of writing itself. Having access to parallel texts has been essential in the deciphering of ancient scripts and to our understanding of ancient civilisations. Moreover, even the earliest parallel and aligned texts had applied purposes similar, albeit not identical, to those we see today, in the sense that apprentice scribes copied and studied such texts as part of their scribal education.

Moving closer to the present time, we have seen that language engineers (e.g. Hofland), translation scholars (e.g. Harris) and contrastivists (e.g. Johansson) have contributed to the development of modern parallel corpora, which are important resources in the applied fields of language teaching and translation studies. In turn, this has led to the renewed interest in contrastive analysis we have witnessed since the mid-1990s.

Primary sources

Holt, Anne. (2004). *Det som aldri skjer*. Oslo: Pantagruel Forlag, 125-164.

Danish translation by Ilse M. Hagaard. (2005). *Det som aldrig sker*. København: Gyldendal.
English translation by Kari Dickson. (2007). *The final murder*. London: Sphere.
French translation by N/A. (2008). *Cela n'arrive jamais*. Paris: Plon.
German translation by Gabriele Haefs. (2007). *Was niemals geschah*. München/Zürich: Piper.
Portuguese translation by N/A. (2008). *Crepúsculo em Oslo*. Lisboa: Quinto selo.
Spanish translation by Cristina Gómez Baggethun. (2008). *Crepúsculo en Oslo*. Barcelona: Roca Editorial.
Swedish translation by Maj Sjövall. (2005). *Det som aldrig sker*. Stockholm: Pirat.

Secondary sources

Brown, P., J. Lai, and R. Mercer. (1991). Aligning sentences in parallel corpora. *Proceedings of the 29th annual meeting of the Association for Computational Linguistics*, *(ACL),* Berkeley, 169-176.

Ebeling, E. (1919). *Keilschrifttexte aus Assur religiösen Inhalts. Erster Band*. Leipzig: J.C. Hinrichs'sche Buchhandlung.

Gale, W.A. and K.W. Church (1991). A program for aligning sentences in bilingual corpora. *Proceedings of the 29th annual meeting of the Association for Computational Linguistics (ACL),* Berkeley, 177-184.

Gale, W.A. and K.W. Church, (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics, 19:*3, 75-102.

Hareide, L. and K. Hofland. (2012). Compiling a Norwegian-Spanish parallel corpus: Methods and challenges. In M.P Oakes & M. Ji (eds.)*. Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research*. Amsterdam/ Philadelphia: John Benjamins Publishing Company, 75-113.

Harris, B. (1988). Bitexts: A new concept in translation theory. *Language Monthly, 54,* 8-10.

Hofland, K. (1996). A program for aligning English and Norwegian sentences. In G. Perissinotto (eds.) *Research in Humanities Computing 5. Selected Papers from the ACH/ALLC Conference, University of California, Santa Barbara, August 1995*. Oxford: Clarendon Press, 165-178.

Hofland, K. and S. Johansson. (1982). *Word frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

Hofland, K. and S. Johansson. (1998). The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In S. Johansson & S. Oksefjell (eds.), 87-100.

Inácio, S. and D. Santos. (2008). Documentação da anotação morfossintáctica da parte portuguesa do COMPARA. http://www.linguateca.pt/COMPARA/DocAnotacaoPortCOMPARA.pdf [accessed 2 September 2012]

Izquierdo, M., K. Hofland, and Ø. Reigem. (2008). The ACTRES parallel corpus: An English-Spanish translation corpus, *Corpora*, 3:1, 31-41.

Johansson, S. (1998). On the role of corpora in cross-linguistic research. In S. Johansson & S. Oksefjell (eds.), 3-24.

Johansson, S. (2007). *Seeing through multilingual corpora: On the use of corpora in contrastive studies.* Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Johansson, S. (2011). Between Scylla and Charybdis. On individual variation in translation. *Languages in Contrast*, 11:1, 3-19.

Johansson, S. and J. Ebeling. (1996). Exploring the English-Norwegian parallel corpus. In C.E. Percy, C.F. Meyer and I. Lancashire (eds.). *Synchronic corpus linguistics*. *Papers from the sixteenth international conference on English research on computerized corpora (ICAME 16)*. Amsterdam: Rodopi, 3-15.

Johansson, S., J. Ebeling, and K. Hofland. (1996). Coding and aligning the English-Norwegian Parallel Corpus. In K. Aijmer, B. Altenberg, and M. Johansson (eds.). *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies*. *Lund 4-5 March 1994*. Lund: Lund University Press, 87-112.

Johansson, S., J. Ebeling, and S. Oksefjell. (1999/2001). *The English-Norwegian Parallel Corpus: Manual*. Institutt for britiske og amerikanske studier, Universitetet i Oslo.

Johansson, S. and K. Hofland. (1989). *Frequency analysis of English vocabulary and grammar based on the LOB Corpus*, Vols. 1 and 2. Oxford: Clarendon Press.

Johansson, S. and K. Hofland. (1994). Towards an English-Norwegian parallel corpus. In U. Fries, G. Tottie, and P. Schneider (eds.), *Creating and using English language corpora*. Amsterdam/ New York: Rodopi, 25-37.

Johansson, S. and K. Hofland. (2000). The English-Norwegian Parallel Corpus: current work and new directions. In S.P. Botley, A.M. McEnery and A. Wilson (eds.) *Multilingual corpora in teaching and research*. Amsterdam / Atlanta, GA: Rodopi, 134-147.

Johansson, S. and S. Oksefjell (eds.). (1998). *Corpora and cross-linguistic research: Theory, method, and case studies*. Amsterdam/ Atlanta: Rodopi.

Lambert, W.G. (1996). *Babylonian wisdom literature*. Winona Lake, Indiana: Eisenbrauns.

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

Mauranen, A. and P. Kujamäki (eds.). (2004). *Translation Universals – Do they exist*? Amsterdam: Benjamins.

McEnery, T. and A. Hardie. (2011). *Corpus linguistics: Method, theory and practise*. Cambridge: Cambridge University Press.

Ohgama, N and E. Robson. (2010). Scribal schooling in Old Babylonian Kish: the evidence of the Oxford tablets. In H.D. Baker, E. Robson and G. Zólomy (eds.) *Your praise is sweet. A memorial volume for Jeremy Black from students, colleagues and friends*. London: British Institute for the Study of Iraq.

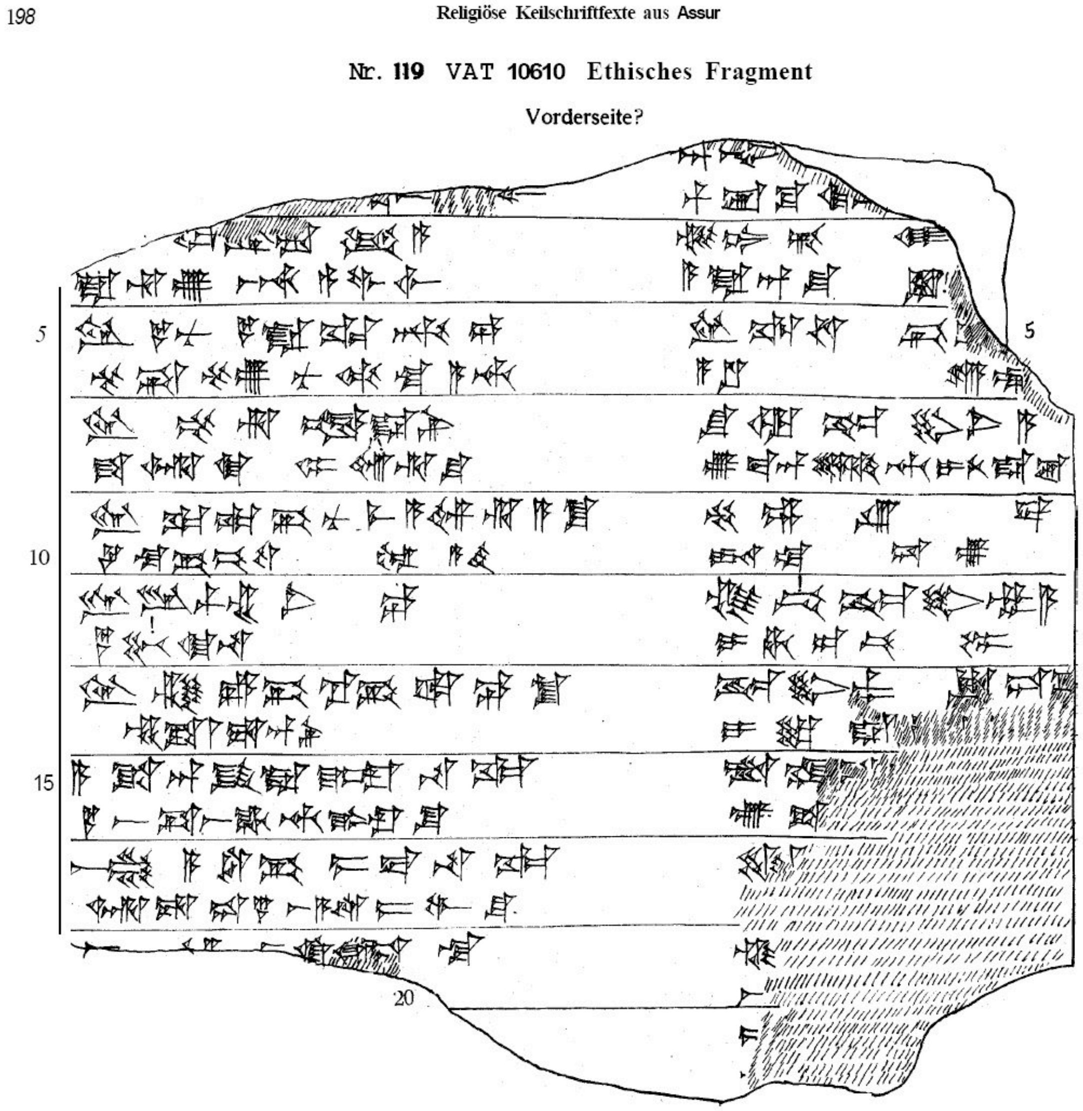
Oksefjell, S. (1999). A description of the English-Norwegian Parallel Corpus: Compilation and further developments. *International Journal of Corpus Linguistics*, 4:2, 197-219.

Santos, D. and S. Oksefjell. (1999). An evaluation of the Translation Corpus Aligner, with special reference to the language pair English-Portuguese. In T. Nordgård (ed.) *NODALIDA'99, Proceedings from the 12th "Nordiske datalingvistikkdager"*. Trondheim: NTNU, 191-205.

Simard, M., G. Foster, and P. Isabelle. (1992). Using Cognates to Align Sentences in Bilingual Corpora. *Proceedings of the fourth international conference on theoretical and methodogical issues in machine translation* (TMI92), (Montreal), 67-81.

Véronis, J. (2000). From the Rosetta stone to the information society: a survey of parallel text processing. In J. Véronis (ed.), *Parallel text processing: Alignment and use of translation corpora,* 1-24. Dordrecht: Kluwer Academic Publishers.

Appendix



198 Religiöse Keilschriftfexte aus Assur

Nr. 119 VAT 10610 Ethisches Fragment

Vorderseite?

Figure A1. Drawing of a clay tablet (A hymn to Ninurta with ethical instructions (Ebeling 1919)