DE GRUYTER OPEN

ICAME Journal, Volume 41, 2017, DOI: 10.1515/icame-2017-0002

A functional comparison of recurrent word-combinations in English original vs. translated texts

Signe Oksefjell Ebeling and Jarle Ebeling, University of Oslo

Abstract

The study explores the potential of quantitative methods to shed light on how texts originally written in English (EO) and texts translated into English (ET) from Norwegian cluster in terms of functional classes. The object of study are sequences of three words (3-grams), classified into 15 functional categories. The investigation establishes that EO and ET do not differ significantly in half of the categories. As for the categories that do differ, two (Comparison and Spatial) are investigated in more detail, uncovering that the more frequent use of Comparison and Spatial 3-grams in ET is most likely a result of source language shining through. The findings are important in the context of both descriptive translation studies and translation-based contrastive studies. With regard to the former, the current study shows that, in many cases, ET does not seem to constitute a 'third code' at the level of 3-gram functions, since the same functions are equally attested in EO. As far as contrastive studies are concerned, the investigation reveals few, if any, lexico-grammatical differences between EO and ET that overturn the belief that translations are a good tertium comparationis *when comparing and contrasting language systems.*

1 Introduction and aims

Corpus-inspired linguistic research has shown that language users to a large extent "build up discourse by means of prepatterned expressions of various kinds" (Altenberg 1993: 17). Studies of how such expressions manifest themselves in different types of language (output) have, among other things, focused on similarities and differences between:

- Text types/registers (e.g. Biber *et al*. 1999; Stubbs and Barth 2003);
- Disciplines (e.g. Cortes 2004; Hyland 2008);
- Learner vs. native-speaker language production (e.g. De Cock 2004; Ädel and Erman 2012);

- Languages (e.g. Cortes 2008; Granger 2014; Ebeling and Ebeling forthc.);
- Original vs. translated text in the same language or across languages (e.g. Baker 2004; Xiao 2011; Ebeling and Ebeling 2013).

This investigation latches on to the last type of study with the aim of shedding more light on how words cluster in English original (EO) texts as compared to English translated (ET) texts. It will thus feed into the discussion of the pros and cons of using translations as a basis for Contrastive Analysis.

Translation data from parallel corpora have been used in Contrastive Analysis for more than two decades despite the fact that concerns have been expressed regarding the use of translations as evidence of language phenomena (see e.g. Teubert 1996 and Mauranen 1998). Two issues in particular are often raised:

- translations distort the target language because of influence from the source language;
- translated language is different from original language; i.e. translation is regarded as a 'Third Code' (Frawley 1984).

On the basis of recurrent word-combinations in the form of 3-grams extracted from a corpus of English original fiction texts and a corpus of English fiction texts translated from Norwegian, we investigate the validity of the second point.[1] Preliminary observations suggest that the original and translated texts seem to cluster in similar ways; e.g. the top 20 3-grams in English originals vs. English translations have an overlap of close to 100 per cent, although some differences can be noted in the ranking of the word-combinations. Our starting point is thus that the two varieties of English (EO / ET) exhibit the same type and number of 3-grams.[2] The aim of this study is to outline and analyse the nature of such combinations in English originals vs. English translations in terms of their function and frequency. More specifically, the study explores to what extent EO and ET 3-grams differ functionally. By taking a statistical approach, we will be able to establish whether variety has a significant effect on the function of 3-grams or not. Another aim of the study is to outline a functional taxonomy and method by which such combinations in the two varieties of English can be adequately compared and analysed. The study can be said to be exploratory and experimental in the sense that it tests the potential of using quantitative methods to steer the qualitative research in interesting and meaningful directions. The following quotation from Firth (1957) underlines the relevance of studying translations and of comparing translations with non-translated texts to uncover how 'fresh ideas' are clothed:

> Do we really know how we translate or what we translate? What is the 'interlingua'? Are we to accept 'naked ideas' as the means of crossing from one language to another? Are these ideas clothed first in Chinese and afterwards in English? Or does the Chinese clothe a collection of naked ideas from which only a selection may accept raiment? And do fresh ideas come in with the English raiment? (Firth 1957: 27)

The many questions posed, and challenges set, by Firth have received a lot of attention within corpus-based contrastive linguistics and descriptive translation studies in recent years. With reference to Firth, this study investigates whether there exist linguistic clues as to how 'ideas' formed in Norwegian and translated into English tally with 'ideas' originally formed in English.[3]

From the perspective of Contrastive Analysis based on original and translated texts, the translations act as the common ground (*tertium comparationis*) upon which hypotheses about similarities and differences between two or more languages can be tested. Translations are considered a good *tertium comparationis*, since it is believed that translators strive to keep and convey not only the ideational and interpersonal functions of the original, but also the stylistic features when translating fiction.

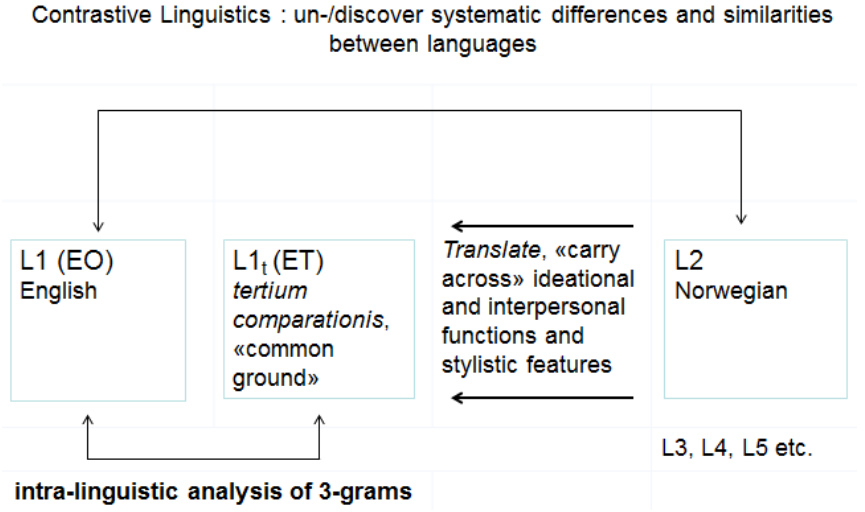

Figure 1: Schematic layout of the procedure of doing contrastive analysis based on translation corpora

Figure 1 shows how functions, in the Hallidayan sense (cf. James 1980: 178), and stylistic features are 'carried across' from one language (and culture) to another when a text is being translated. Moreover, it shows the importance of studying the result, i.e. the translation, to be able to pin down those features that, in the words of Teich (2003), are due to source language (SL) shining through or target language (TL) normalization, or indeed, general features stemming from the translation process itself. Depending on your area of interest, Translation Studies, Contrastive Linguistics, Social Anthropology, etc., the features you are able to recognise as different (or similar) to features of the target language will feed into your research in different ways. For the contrastivist, interested in systemic and systematic differences between two or more languages, the features must be accounted for in such a way as not to impinge on the result of the contrastive analysis.

The outline of this article is as follows. We start, in Section 2, by introducing some previous research focusing on original and translated texts in the same language, mainly concerned with sequences of words. Section 3 outlines the material and method used, including issues to do with data extraction (3.2) and the functional classification (3.3). Further issues to do with the normalization of 3-gram frequencies are discussed in Section 3.4, while an introduction to the statistical approach adopted with some quantitative findings are offered in Section 3.5. A discussion of the more qualitative findings is offered in Section 4, followed by the conclusion in Section 5.

2 Previous research

A large number of studies that in some way or other are tangent to, or partially overlap with, the current investigation have been carried out over the past 20 years or so.[4] Some of these are relevant because of their methodological framework of how to classify sequences of words in text, while others are relevant because of their concern with original vs. translated text in the same language. As we cannot do justice to all this previous work, we will confine ourselves to a handful of studies of the latter type that bear most relevance to the present one. As to the taxonomy used for the classification of 3-grams, which will be presented in Section 4.3, it is mainly inspired by Altenberg (1998), Moon (1998) and Biber *et al*. (2004).[5]

Baker (2004, 2007) and Xiao (2011) investigate recurrent multi-word sequences in original vs. translated texts in the same language, albeit with slightly different agendas from ours. Nevertheless, these studies are relevant to the current one, particularly with regard to initial data extraction and the focus

on intra-linguistic issues, viz. translated vs. non-translated English in the case of Baker, and translated vs. non-translated Chinese in the case of Xiao.

One of the case studies presented in Xiao (2011) is concerned with the comparison of the frequency of word clusters in translational and native Chinese. The comparisons reported are, however, confined to surveying main tendencies only, in terms of overall frequencies of 2-, 3- and 4-word clusters, overall frequencies of high-frequency word clusters and the overall coverage of 2- and 3-word clusters. Although we aim to make some observations regarding overall tendencies in the use of 3-grams in our material, our main concern is, as previously stated, the functional nature of the word sequences in English original vs. translated texts.

Restrictions apply to the sequences studied by both Baker (2004, 2007) and Xiao (2011); Baker (2004), for example, excludes "combinations of words that are not recognizable as fixed or semi-fixed lexical phrases" (2004: 174), while Xiao (2011) moves from the general overview outlined above to investigate sequences representing reformulation markers. Moreover, Xiao, in particular, relates his findings to the concept of translation universals. In some sense, then, their focus is narrower than what is aimed at here, where we wish to report general tendencies for a host of word sequences, amounting to 1,400-1,500 3-gram types in each of our sub-corpora. On the other hand, Xiao's (2011) study has a broader focus in that it includes a comparison with source language (English) texts, in order to establish its potential influence on the Chinese translations. Xiao's exploitation of corpus resources resembles that of Teich (2003), particularly as regards points (i) and (iii) below.

In her comprehensive investigation of cross-linguistic variation in system and text, Teich (2003) compares features of English and German from three different perspectives: (i) original vs. original texts in the two languages; (ii) original vs. translated texts in the two languages; (iii) original vs. translated texts in the same languages. The third type of comparison poses similar questions to those at the heart of this investigation, e.g. "What are the typical lexico-grammatical features of […] an English translation from German" (2003: 2), and how do these compare with features of texts originally written in English within the same register? Teich's method differs from the current "knowledge-free method" (Baroni and Bernardini 2003: 85) in taking "previously attested register features" as the starting point to investigate translations and comparable texts (Teich 2003: 229); her focus is on features such as transitivity, voice and NP complexity rather than on strictly functional categories. Moreover, although Teich is mostly preoccupied with developing a sound method and theoretical framework for multilingual studies, catering for typological, translation and

contrastive studies, her analysis reveals that "translations indeed exhibit a mixture of SL shining through and TL normalization" (2003: 222).[6]

With reference to Teich's point (i), it should be mentioned that a comparison of English originals vs. Norwegian originals, following similar methodological steps as in the current study, will be conducted in a separate study.[7] It is believed that results from such an inter-linguistic comparison may feed directly into this intra-linguistic one, as we will be able to point to (potential) SL effects (from Norwegian) in the English translations, which in turn may explain (potential) deviations from the use of 3-grams in English originals. In fact, Teich points out that "[h]aving established the relation between English and German cross-linguistically comparable texts, it can be used as a basis for working on the second and third" types of comparison (2003: 2).

Inspired by Baker (2004) and Biber and others (e.g. Biber *et al*. 1999; Biber *et al*. 2004), Lee (2013) analyses lexical bundles with the aim of uncovering characteristic features of translated Korean as compared to non-translated Korean. Lee outlines the structural and functional taxonomies proposed by Biber *et al*. (2004) and explicitly relates the extracted bundles to Biber's classification scheme by stating that temporal and spatial phrases can be subsumed under the category referential bundles (Lee 2013: 383); and further "the bundles on the list belong to either of the other two categories – stance and discourse organizing bundles" (Lee 2013: 384). Thus, the relevance of Lee's study, with such functional categories at its core, should be obvious (Section 3.4). However, the case studies in Lee (2013) do in fact bear more resemblance to Baker's (2004) study in that the initial bundle lists only serve as a starting point for narrowing down the object of study to individual bundles that are "likely to point to important differences between translated and non-translated texts" (Lee 2013: 383).

This section has served to illustrate that translated vs. non-translated texts in the same language have been studied using techniques similar to those applied in the present study, but that to date, no large-scale comparisons of the type we have seen for text-types/registers and disciplines have emerged. Incidentally, Lee (2013: 393) addresses this issue, and compares the method proposed for translation research with "the conventional methods used by corpus linguists and stylisticians researching phraseological variation across genres and registers" (Lee 2013: 392–393). Lee concludes that comparisons of overall distributions "between the corpora under examination" are "irrelevant to our interest in uncovering features characteristic of translated texts, independent of source texts." The reason for this is that many bundles "are still traceable to the source texts, meaning that they mirror the properties of the source texts" (Lee 2013:

393). This view echoes to some extent Mauranen (2007), who stresses the need "to consider factors other than overall tendencies of a very general kind", particularly in view of her observation that "[d]etailed analyses tend to show that the behavior of different linguistic items is not identical" (Mauranen 2007: 40). These quotes and the observations reported above by Teich remind us that it is important to bear in mind that the piece of research presented next is limited in that the translations we investigate have only one source language, and that we only look at translations into English.

Despite the scepticism towards using overall distribution, or tendencies, as points of departure when comparing original and translated texts, and the ongoing discussion of the existence of translation universals within translation studies,[8] it is important for contrastive linguists in particular to unearth and investigate potential differences between texts originally written in a language and texts translated into that language, if such data are to be used as a basis for our contrastive research also in the future. We hope the method outlined in the following sections can go some way towards identifying the nature of these differences.

3 Material and method

3.1 The corpus

The material used in this study is culled from the English-Norwegian Parallel Corpus+ (ENPC+). It is a balanced and bidirectional translation corpus containing 39 fictional texts originally written in English (EO) and 39 in Norwegian (NO) and their translations (NT, ET). The ENPC+ contains extracts of books of between 10,000 and 17,000 words (Johansson *et al*. 1999/2001), as well as some full-length novels of between 52,000–211,000 words (8 English original texts and 9 Norwegian original texts). Each part of the corpus amounts to around 1.3 million (EO, NO, NT) and 1.4 million (ET) words (see Ebeling and Ebeling 2013: 84ff for more detailed information about the ENPC+). Thus, the four subcorpora are comparable in terms of size, both with regard to number of texts and number of running words. They are also comparable in the sense that they contain texts that can be broadly defined as contemporary fiction (1980s–2012). For the purpose of this study only the EO and ET sub-corpora of the ENPC+ will be used.

3.2 Data delimitation and extraction

As our corpus is relatively small, we have chosen to focus on 3-grams to get a large, but manageable, set of sequences to analyse, well aware of the challenges

we may encounter in the functional classification of such short sequences, and the fact that two 3-gram sequences can be seen to belong in a sense to the same 4-gram. The 3-grams were extracted using AntConc,[9] in which some changes to the default settings were made (for a more detailed description, see Ebeling and Ebeling forthc.).

In our investigation we also try to ensure that our results are not due to idiosyncratic uses by the individual authors or translators; we therefore require the 3-grams to occur in at least 25 per cent of the texts, i.e. in 10 of the 39 texts. We also introduced an additional, and quite conservative, threshold requiring each 3-gram to occur with a frequency of at least 20 per million words (pwm),[10] i.e. 26 and 28 times in EO and ET, respectively. We refer to these two conditions, distribution and recurrence, collectively as **the threshold**.

The method of extraction gave us two comparable lists of 3-gram types (i.e. different 3-grams), one for the EO texts and one for the ET texts, amounting to 1,408 and 1,468 3-gram types, yielding 83,827 vs. 87,878 3-gram tokens, respectively. At this point, it can be observed that there is a statistically significant difference between the sub-corpora when we compare the token counts, but not when we compare the type counts. In both cases we used the total number of 3-grams as baseline (EO: 1,110,300 vs. ET: 1,119,699) where the `prop.test` in R returns the following *p*-values for types and tokens, respectively: p=0.3817; p<0.0001; df=1.

3.3 Classification of data

For the functional classification of the 3-grams we decided on a mixed taxonomy, inspired by Altenberg (1998), Moon (1998) and Biber *et al*. (2004). We draw mainly on Altenberg (1998), but some elements are also recognisable from Moon's (1998) and Biber *et al*.'s (2004) taxonomies. We operate with four main functional categories, viz. Evaluative, Informational, Modalizing and Organizational. The Informational category is further divided into 12 subcategories. In the list below, an overview is given of the altogether 15 categories; the 12 Informational categories are left unmarked, while the non-Informational ones are marked in bold. The number following the examples given for each category shows the approximate frequency of 3-gram types.

- Comparison (*as good as, as if to, looked like a*) <= 25
- Contingency (*because it was, if he 'd, why did you*) 50–100
- **Evaluative** (*'s a good, i 'm sure, just do n't*) EO: 51, ET: 30
- Existential (*and there 's, there were no*) <= 25
- Fragment (*a sense of, the door and, to go on*) 50–100

- **Modalizing** (*'ll tell you, but he could, seemed to be*) > 100
- **Organizational** (*all the same, in any case*) <= 25
- Process (*in a way, the way you*) <= 25
- Quantifying and Intensifying (*a glass of, more or less, lot of time*) 50–100
- Reporting (*he said and, no he said*) <= 25
- Respect (*apart from the*)[11] 1
- Rhematic (*'s not a, he told me, to give him*) > 100
- Spatial (*across the table, back in the, to be there*) > 100
- Temporal (*a few days, at the moment, he 'd never*) 50–100
- Thematic stem (*and i 'm, but he had, what 's happened*) > 100

The category Comparison is rather small and many of the 3-grams start with the sequence *as if*. Contingency 3-grams express a condition, reason, cause or concession. Evaluative 3-grams are similar to the Modalizing ones, but typically contain an evaluative adjective or adverb instead of a verb. Existential requires existential *there* as part of the 3-gram. Fragments typically consist of noun phrase(s) (fragments) that could be either thematic or rhematic. Some verb phrase(s) (fragments) are also found in this category. Modalizing 3-grams contain verbs that are either identifiable as modal auxiliaries or other items, mostly other verbs, expressing attitude, possibility/probability or certainty towards a proposition, e.g. *know, think, want to, perhaps* and *seem*. Organizational 3-grams are represented by 3-grams that are clearly recognizable as text structuring devices, e.g. connectors. Process is represented by manner and means expressions. Quantifying and Intensifying 3-grams are included in the same category, since it is often difficult to categorically say whether a 3-gram is quantifying or intensifying something. Reporting 3-grams include a reporting verb, usually *said*. Rhematic 3-gram types typically include a verb followed by (part of) a noun phrase (i.e. the beginning of an object or complement/ predicative). The categories Spatial and Temporal include a spatial or temporal element referring to space or time, typically in the form of prepositions and adverbs. Finally, Thematic stems "consist of subject and verb (plus any preceding thematic elements) but lack a rhematic post-verbal element" (Altenberg 1998: 111).

To determine the function of a 3-gram type, a few simple guidelines were followed. First of all, a 3-gram type cannot have dual membership. This means, for instance, that Modalizing is chosen in all cases where a clearly modal element is present (typically a modal auxiliary verb). The only exception here is when *if* is part of a 3-gram introducing a conditional clause; in these cases, Contingency trumps Modalizing (e.g. *if you can*). Contingency (*if*-clause) is also

chosen over other categories, such as Existential (*if there 's*). The category Thematic Stem is chosen when a stem is not Modalizing, Spatial, Temporal etc. in nature; in other words, it could be characterized as a neutral, but Informational stem. Similar rules were followed for Fragment.

Relatively few conflicts between categories were noted. However, in these cases and in truly ambiguous cases, the membership of a 3-gram type is based on its most common use in the corpus; i.e. the context was consulted to decide its category membership.

3.4 Normalization of frequency counts

Following the classification of the 3-gram types that met the threshold in each of the two sub-corpora, the token counts for each functional category were registered.[12] The total number of 3-grams for each text was also counted and inserted into our matrix. A snapshot of these counts is given in Table 1, represented by counts from two texts from EO and two from ET, for two of the functional categories, viz. Fragment and Modalizing.

Table 1: Token counts and normalized frequencies

	EO		ET	
	MoAl1E[13]	MW1E	JoNe1TE	JW1TE
# of 3-grams	63,068	8,878	113,843	11,326
Fragment tokens	180	15	524	50
Modalizing tokens	1,058	158	1,774	104
Normalized frequencies	Fragment / Modalizing		Fragment / Modalizing	
Tokens / 3-grams * 1,000	2.85 / 16.78	1.69 / 17.8	4.6 / 15.58	4.41 / 9.18

The normalized frequencies in Table 1 show the number of 3-gram tokens that meet the threshold per 1,000 of all 3-gram tokens in a text. These numbers form the basis for the statistical tests that are run on all 39 texts in the EO and ET corpora. For instance, for text MW1E, the number of Modalizing tokens is 158. This number is divided by 8,878 (= # of 3-grams), resulting in a normalized frequency of 17.8 per 1,000 3-grams.

During the initial extraction of the 3-gram types we noticed that all the types attested in the EO texts that reached our threshold were in fact attested in the ET texts and vice versa, even though they did not reach the threshold in the respective sub-corpora; either they did not occur at least 20 times pmw or were not

attested in at least 25 per cent of the texts. It was therefore decided that the counts should be evened out before the analysis; thus, in the following the token counts for the 3-gram types that did not initially reach the threshold in either EO or ET are also included, giving us the same amount of 3-gram types for both EO and ET. Table 2 shows five 3-gram types of the Temporal category. One of the 3-grams reaches the threshold in both corpora, viz. *for a while*, while the remaining only reach the threshold in either EO or ET. In addition to not reaching the threshold of 20 occurrences pmw (26 times in EO and 28 times in ET), two 3-grams are found in only four texts: *for many years* in EO and *for the day* in ET. *For as long* and *for so long* reach the distribution threshold in EO but not the frequency threshold.

Table 2: Added token counts for 3-gram types that initially did not reach the threshold

3-gram	EO freq.	EO distr.	ET freq.	ET distr.
for a while	152	>= 10	188	>= 10
for as long	20	13	28	>= 10
for many years	5	4	34	>= 10
for so long	21	11	30	>= 10
for the day	29	>= 10	5	4

The shaded cells in Table 2 show the number of tokens that were added for 3-gram types that did not initially reach the thresholds for either EO or ET. For, e.g., *for so long* in EO, only five more attested occurrences would have meant that the Temporal category would have included 26 more instances in the EO corpus distributed over the 39 texts.

3.5 Quantitative comparison of the functional categories

We used an independent, two-tailed *t*-test with Welch's correction as implemented in R to compare the normalized frequencies of the functional categories in EO vs. ET.[14] The normalized frequency for 3-gram tokens in each functional category is, as mentioned above, measured against the total number of 3-gram tokens (in each text). At a confidence level of 95%, the *p*-values we obtain either reject (significant *p*-value) or fail to reject (non-significant *p*-value) our hypothesis that the two varieties use functional categories of 3-grams with a similar frequency.

It should be noted at this point that we primarily see the *p*-values stemming from the *t*-test as indicators pointing to interesting avenues for further qualitative, linguistic study; for example, if we get a significant *p*-value, can further qualitative inquiry uncover reasons for this? Has it to do with source language shining through, target language normalization, or could it be due to something else?

Table 3 shows the *p*-values for the 14 categories we have identified (excluding Respect). Interestingly, a statistically non-significant result is found for half of the categories (shaded cells), meaning that we cannot rule out chance as a factor, thus suggesting that EO and ET behave similarly at this functional level of analysis for these categories.[15]

Table 3: *p*-values calculated for each functional category

Category	*t*-score$_{(df)}$	*p*-value	Favoured in
Comparison	$-3.13_{(52.86)}$	$p = 0.002$	ET
Contingency	$0.41_{(75.709)}$	$p = 0.679$	--
Evaluative	$1.26_{(72.225)}$	$p = 0.210$	
Existential	$0.51_{(75.98)}$	$p = 0.605$	--
Fragment	$-3.40_{(72.586)}$	$p = 0.001$	ET
Modalizing	$1.01_{(74.942)}$	$p = 0.313$	--
Organizational	$-3.15_{(58.596)}$	$p = 0.002$	ET
Process	$-2.45_{(75.374)}$	$p = 0.016$	ET
Quantifying/Intensifying	$-1.04_{(75.048)}$	$p = 0.299$	--
Reporting	$2.85_{(72.22)}$	$p = 0.005$	EO
Rhematic	$0.64_{(75.732)}$	$p = 0.522$	--
Spatial	$-3.91_{(74.638)}$	$p < 0.001$	ET
Temporal	$-4.13_{(74.824)}$	$p < 0.001$	ET
Thematic stem	$0.01_{(73.245)}$	$p = 0.991$	--

The right-most column in Table 3 shows that in all but one of the categories showing a statistically significant result, viz. Reporting, the functional category is more used in ET than in EO. In the following section, we will take a closer look at two of the six categories in which this is the case, in an attempt to uncover in what way and why ET should boast a higher use of these 3-gram functions.

4 Qualitative findings and discussion

In this section some attention will be given to the actual 3-grams and frequencies that give rise to the significant *p*-values for the categories Comparison and Spatial. The 3-gram token counts will be sorted by the difference in the number of tokens between EO and ET. It should be noted that the numbers reported below are raw frequencies, which means that the differences are inflated by the relative difference in size between the two sub-corpora, i.e. approx. 7.6 per cent, since the ET sub-corpus contains 100,000 words more than the EO sub-corpus.

4.1 Comparison

Comparison is a rather small category with 19 3-gram types in EO and 25 in ET. When combining the two type lists we get 28 3-gram types. Table 4 tells us that out of the top 15 3-grams that show the greatest difference in numbers of tokens, 13 belong to ET.

Table 4: Comparison sorted by difference in number of tokens (raw frequencies)

Variety	Diff.	EO	Freq.	Dist.	ET	Freq.	Dist.
ET	92	it was as	47		it was as	139	
ET	75	as if he	152		as if he	227	
ET	69	was as if	35		was as if	104	
ET	67	as if it	65		as if it	132	
ET	56	as though he	18	8	as though he	74	
ET	53	as if the	53		as if the	106	
ET	42	as if to	26		as if to	68	
ET	37	as if they	52		as if they	89	
ET	36	the same as	23	9	the same as	59	
ET	35	as if i	39		as if i	74	
ET	30	as if she	97		as if she	127	

EO	25	than he had	33		than he had	8	5
ET	23	as if someone	12	8	as if someone	35	
ET	22	was just as	13	7	was just as	35	
EO	21	as well as	80		as well as	59	

The difference (Diff.) between ET and EO for the 3-gram *it was as* is 92, for *as if he* 75 and so on. The actual frequencies underlying the difference are listed in the Freq. columns. The distribution (Dist.) columns indicate whether the particular 3-gram was not initially (before the top-up) among the 3-grams that reached the threshold for that variety. The 3-gram *as though he*, for instance, only occurs in eight of the EO texts with a frequency of 18 per million words (required 26). These clear differences in number of tokens between ET and EO give rise to the significant *p*-value.

The focus of this paper is on overall tendencies in the 15 functional categories and not on actual and detailed linguistic and/or cultural differences between EO and ET gleaned from the texts. We will therefore refrain from going into, and discuss, every difference in frequency of every individual 3-gram. However, with regard to the huge difference in the case of *it was as*, we observe, when looking at the concordance lines for this sequence in the corpora, that this is the initial part of two 4-grams in the ET texts in particular: *it was as if* and *it was as though*, both having, for the most part, the very frequent Norwegian 4-gram *det var som om* as their source.[16] When we look at the sequence *it was as* in the EO sub-corpus, *it was as if* is fairly frequent, but *it was as though* does not occur at all. It seems, then, that (at least) two tendencies come together and create this huge difference; one is source language shining through, in the form a very frequent Norwegian 4-gram, *det var som om* 'it was as if', and the use of *it was as though* on the part of the translators as an alternative to *it was as if*, a use not reflected in the English original texts. Both translation solutions lead to the frequent use of the 3-gram *it was as* in the ET sub-corpus.

4.2 Spatial[17]

Table 5 shows the top 15 Spatial 3-grams sorted by the difference in raw frequencies between EO and ET, all of which show a greater number of occurrences in ET.

Table 5: Spatial sorted by difference in number of tokens (raw frequencies)

Variety	Diff.	EO	Freq.	ET	Freq.
ET	351	in front of	237	in front of	588
ET	133	over to the	61	over to the	194
ET	127	on the other	96	on the other	223
ET	122	on the floor	117	on the floor	239
ET	113	front of the	84	front of the	197
ET	112	in the middle	130	in the middle	242
ET	111	the middle of	121	the middle of	232
ET	106	out into the	39	out into the	145
ET	99	middle of the	72	middle of the	171
ET	97	down on the	81	down on the	178
ET	82	up to the	85	up to the	167
ET	80	down in the	22	down in the	102
ET	80	out of the	443	out of the	523
ET	77	down to the	77	down to the	154
ET	76	of the window	28	of the window	104

Some of these differences, e.g. *in front of*, are striking. Why should there be 351 more occurrences of this 3-gram type in two seemingly similar corpora? If we once more take a look behind the scenes, we notice that *in front of* in the ET subcorpus has the Norwegian word *foran* as their main source. In fact more than 450 of the 588 occurrences have this one-word correspondent. This leads us to suspect that the translators use *in front of* as a default translation when faced with *foran* and do not consider other possible translations (and why should they?), e.g. *before* or *ahead* (*of*), which would work equally well on many occasions. Additionally, there may be some interesting cultural differences regarding the use of spatial expressions in fiction in the two languages, leading to the source language shining through in ET; e.g. a felt need to anchor discourse in space is more common in Norwegian than in English. However, to establish this with more certainty, a more thorough investigation of spatial expressions in English and Norwegian original texts in general is called for.

4.3 Summary of findings

The discussion of the findings for the categories Comparison and Spatial has revealed that the translation sub-corpus, ET, has more token occurrences of the frequent 3-gram types than EO, even if we take into account that ET is slightly larger than EO. This is in line with other studies of original and translated texts and corroborates the findings from these studies that translations show TL normalization and SL shining through (Laviosa 2002; Teich 2003; Xiao 2011), and that these effects together result in a higher token frequency for ET, which, in our case, leads to significant differences reported by the statistical tests. Moreover, for 3-grams in these two categories, "the general rule that frequent items occur even more frequently in translation" (Mauranen 2000: 10) seems to be at play.

More importantly, however, the findings have shown that these differences do not affect all functional categories, at least not in translation from Norwegian. It follows from this that a careful classification of n-grams is useful, if not a pre-requisite, if the purpose of study is to make claims about similarities and differences of original and translated texts.

5 Conclusion

In this study we have presumed, without problematizing it to any great extent, that the ET sub-corpus contains textual equivalents of their sources. It follows from this that we have taken for granted that each translation matches its source ideationally and interpersonally in the Hallidayan sense. Moreover, we believe that the translators strive to keep the style of the source in terms of e.g. length and composition of sentences, the use of direct and indirect speech and thought, the employment of (grammatical) metaphor, etc. as far as this is culturally and grammatically (textually) possible when coding a literary text in a different language. This fundamental premise has made it possible to compare the translations with comparable texts originally written in English at the level of function as defined here. Note, however, that we have not assumed that this can, or will, not lead to instances of non-equivalence at some lower level of analysis, e.g. at the level of the clause, phrase or word (see Halliday 2001 for a discussion of these points).

The conservative threshold employed in the current study hides many interesting facts about the differences and similarities between original and translated English. One difference we noticed very early on, but which is not picked up by the procedure because of, among other things, the way the material was classified into functional categories, is the use of contracted forms. There

appears to be some reluctance among the translators to use contracted forms, but further study, along the lines of Olohan (2003), is needed to establish the exact nature of this phenomenon in our data. Another point we noticed was the way cultural and geographical differences shone through, even with the relatively limited amount of data and conservative threshold applied, e.g. in the form of the frequencies of the grams *cup of tea* (EO: 26 / ET: 12) vs. *cup of coffee* (EO: 11 / ET: 33) and the translators' frequent use of *in the mountains* (EO: 4 / ET: 39) and *in the valley* (EO: 3 / ET: 20).

In another study it would be interesting to classify the 3-grams according to other criteria than functional ones, to see if other kinds of linguistic and cultural differences between English and Norwegian (and other languages) could be unearthed. Finally, and most obviously, studies incorporating translated English fiction from other languages would greatly enhance and extend the generality of studies of this kind.

The method outlined and tested in this paper addresses the questions of whether translations can be used as a *tertium comparationis* (common ground) when doing contrastive analysis and in what ways translations are different from, and similar to, non-translated texts in the same language. The first question can of course not be answered without a good knowledge of the second.

To counter the criticism levelled against using "overall tendencies of a very general kind" (Mauranen 2007: 40) in the study of features of translated texts, we classified our overall tendencies', in the form of frequent 3-grams, into 15 functional categories. The method employed also took into account distribution and recurrence as important factors to make provisions for idiosyncrasies that arise when texts of different size and composition are compiled to make a corpus. Finally, we used a statistical test as a litmus test in order to decide (quantitatively) which functional categories to investigate qualitatively. The results of the study reveal few, if any, lexico-grammatical differences between EO and ET that overturn our belief that translations are a good *tertium comparationis* when comparing and contrasting language systems. This does not, of course, mean that one must not take care when using translations, since we get a skewing effect when the source language shines through and when the translators select (unconsciously) a default target language rendering, whatever the reason. The findings are important not only in the context of Contrastive Analysis, but also in the context of Descriptive Translation Studies. The current study shows that, in many cases, ET does not seem to constitute a "third code" at the level of 3-gram functions, since the same functions are equally attested in EO.

Notes

1. A 3-gram is an uninterrupted sequence of three words, regardless of semantic unity, e.g. *by the way*, *had been no*.
2. For lack of a better term, we use 'variety' when referring to EO and ET throughout.
3. It is not altogether clear (to us) what Firth meant by 'idea' in this context, but we interpret it to mean, not only social and cultural artefacts (things) and concepts, but also ways of packaging information, i.e. conveying meaning, and the establishing and maintaining of interpersonal relations in different languages.
4. Altenberg (1998), Baroni and Bernardini (2003), Biber (2006), Biber *et al*. (2003; 2004), Biber and Barbieri (2007), Chen and Baker (2010), Cortes (2008), Ebeling *et al*. (2013), Ebeling and Ebeling (2013), Granger (2014), Hyland (2008), Kermes and Teich (2012), Lee (2013), Mauranen (2000), Moon (1998), Nattinger and DeCarrico (1992), Oakes and Ji (2012), Stubbs and Barth (2003), Teich (2003), Xiao (2010) to mention a few.
5. See Ebeling and Ebeling (forthc.) for an overview and a discussion of the three functional frameworks.
6. 'Translationese' (Gellerstam 1986) and 'translation effect' (Johansson 2007) are sometimes uses synonymously with S(ource)L(anguage) shining through.
7. Ebeling and Ebeling (forthc.).
8. The concept of translation universals is challenged by House (2008: 11): "For the present author, the functional base underlying language use as suggested by Halliday […] are [sic] a prime candidate for universalism in translation. But: these are then not universals of translation per se, or sui generis universals, but simply universals of language also applying to translation".
9. http://www.laurenceanthony.net/software/antconc/
10. This is in line with Biber *et al*.'s (2003: 74, 75) cut-off frequency of 20 times pmw, although later in their article the cut-off frequency seems to have been adjusted to 40 times pmw (p. 78).
11. The Respect category is represented by one 3-gram type only, *apart from the*, and will not be part of the discussion below.
12. The tokens were counted by a Perl script reading each text and comparing every 3-gram in that text with the 3-grams in the type lists.
13. See Ebeling and Ebeling (2013) for an overview of the texts (and text identifiers) included in the EO and ET.

14. R version 3.2.4.
15. Not all of our data are normally distributed, so we also ran the non-parametric Mann-Whitney-Wilcoxon test on the same data and it showed the same tendencies with regard to significant vs. non-significant *p*-values.
16. An alternative to *det var som om* 'it was as if' is *det virket som om* 'it seemed as if', which is also translated into *it was as if*. Thus, there are two Norwegian expressions with similar meaning giving rise to one English correspondent.
17. The Spatial category is also discussed in an inter-language context in Ebeling and Ebeling (forthc.).

References

Ädel, Annelie and Britt Erman. 2012. Recurrent word combinations in academic writing by native speakers and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes* 31: 81–92.

Altenberg, Bengt. 1993. Recurrent word combinations in spoken English. In J.M. D'Arcy (ed.). *Proceedings of the Nordic Conference for English Studies, Reykjavik, 7–8 August 1992,* 17–27. Reykjavik: Publications of the Institute for Foreign Languages, Faculty of Arts, University of Iceland.

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In A.P. Cowie (eds.). *Phraseology: Theory, analysis and applications*, 101–122. Oxford: Oxford University Press.

Baker, Mona. 2004. A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics* 9(2): 167–194.

Baker, Mona. 2007. Patterns of idiomaticity in translated vs. non-translated English. *Belgian Journal of Linguistics* 21: 11–21.

Baroni, Marco and Silvia Bernardini. 2003. A preliminary analysis of collocational difference in monolingual comparable corpora. *UCREL Technical Paper* number 16 (Special issue). In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster University (UK), 28–31 March 2003. http://ucrel.lancs.ac.uk/publications/cl2003/papers/baroni.pdf. Accessed 1 June, 2016.

Biber, Douglas. 2006. *University language. A corpus-based study of spoken and written registers.* Amsterdam: Benjamins.

Biber, Douglas and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263–286.

Biber, Douglas, Susan Conrad and Viviana Cortes. 2003. Lexical bundles in speech and writing: an initial taxonomy. In A. Wilson, P. Rayson and T. McEnery (eds.). *Corpus linguistics by the lune: A festschrift for Geoffrey Leech*, 71–105. Frankfurt: Peter Lang.

Biber, Douglas, Susan Conrad and Viviana Cortes. 2004. 'If you look at…:' Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371–405.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English.* London: Longman.

Cortes, Viviana. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23: 397–323.

Cortes, Viviana. 2008. A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora* 3(1): 43–57.

De Cock, Sylvie. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literature (BELL)* New Series 2: 225–246.

Ebeling, Jarle and Signe Oksefjell Ebeling. 2013. *Patterns in contrast*. Amsterdam: John Benjamins.

Ebeling, Jarle, Signe Oksefjell Ebeling and Hilde Hasselgård. 2013. Using recurrent word-combinations to explore cross-linguistic differences. In K. Aijmer and B. Altenberg (eds.). *Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson*, 177–199. Amsterdam: John Benjamins.

Ebeling, Signe Oksefjell and Jarle Ebeling. Forthc. A cross-linguistic comparison of recurrent word-combinations in a comparable corpus of English and Norwegian fiction. To appear in M. Janebova, E. Lapshinova-Koltunski and M. Martinkova (eds.). *Contrasting English through corpora. Corpus-based contrastive analysis of English and other languages*. Edinburgh: Cambridge Scholars.

Firth, John R. 1957. A synopsis of linguistic theory, 1930–1955. *Studies in linguistic analysis*, 1–32. Oxford: Basil Blackwell.

Frawley, William. 1984. Prolegomenon to a theory of translation. In W. Frawley (ed.). *Translation. Literary, linguistic and philosophical perspectives*, 159–175. Newark: University of Delaware Press.

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

Granger, Sylviane. 2014. A lexical bundle approach to comparing languages: Stems in English and French. In M-A. Lefer and S. Vogeleer (eds.). *Genre- and register-related discourse features in contrast*. Special issue of *Languages in Contrast* 14(1): 58–72.

Halliday, M.A.K. 2001. Towards a theory of good translation. In E. Steiner and C. Yallop (eds.). *Exploring translation and multilingual text production: Beyond content*, 13–18. Berlin: Mouton de Gruyter.

House, Juliane. 2008. Beyond intervention: Universals in translation? *trans-kom* 1(1): 6–19.

Hyland, Ken. 2008. 'As can be seen.' Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4–21.

James, Carl. 1980. *Contrastive analysis*. London: Longman.

Johansson, Stig. 2007. *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Johansson, Stig, Jarle Ebeling and Signe Oksefjell. 1999/2001. *The English-Norwegian Parallel Corpus: Manual.* Department of British and American studies, University of Oslo. http://www.hf.uio.no/ilos/english/services/omc/enpc/ENPCmanual.pdf. Accessed 1 June, 2016.

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

Lee, Changsoo. 2013. Using lexical bundle analysis as discovery tool for corpus-based translation research. *Perspectives* 21(3): 378–395.

Mauranen, Anna. 1998. Form and sense relations as seen through parallel corpora. In W. Teubert, E. Tognini-Bonelli and N. Volz (eds.). *Proceedings of the Third European Seminar "Translation Equivalence". Montecatini Terme, Italy October 16–18, 1997*, 159–173. Germany: The TELRI Association e.V.

Mauranen, Anna. 2000. Strange strings in translated language: A study on corpora. In M. Olohan (ed.). *Intercultural faultlines. Research models in translation studies I: Textual and cognitive aspects*, 119–141. Manchester: St Jerome.

Mauranen, Anna. 2007. Universal tendencies in translation. In G.M. Anderman and M. Rogers (eds.). *Incorporating corpora. The linguist and the translator*, 32–48. Clevedon: Multilingual Matters.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English. A corpus-based approach*. Oxford: Clarendon Press.

Nattinger, James R. and Jeanette S. DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

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

Olohan, Maeve. 2003. How frequent are contractions? A study of contracted forms in the Translational English Corpus. *Target* 15(1): 59–89.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. Accessed 1 June, 2016.

Stubbs, Michael and Isabel Barth. 2003. Using recurrent phrases as text-type discriminators. A quantitative method and some findings. *Functions of Language* 10(1): 61–104.

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

Teubert, Wolfgang. 1996. Comparable or parallel corpora? *International Journal of Lexicography* 9(3): 238–264.

Xiao, Richard. 2010. How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics* 15(1): 5–35.

Xiao, Richard. 2011. Word clusters and reformulation markers in Chinese and English: Implications for translation universal hypotheses. *Languages in Contrast* 11(2): 145–171.