Meta
Journal des traducteurs
Translators' Journal

The function of recurrent word-combinations in English translations from three different languages

Signe Oksefjell Ebeling

Volume 67, numéro 1, avril–mai 2022

Pour de nouvelles méthodes en traductologie quantitative
Exploring New Methods in Quantitative Translation Studies

URI : https://id.erudit.org/iderudit/1092194ar
DOI : https://doi.org/10.7202/1092194ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Ebeling, S. O. (2022). The function of recurrent word-combinations in English translations from three different languages. *Meta*, *67*(1), 143–169. https://doi.org/10.7202/1092194ar

Résumé de l'article

Cet article compare les tendances phraséologiques dans les textes anglais traduits et non traduits en utilisant des séquences de trois mots classés selon leur fonctionnement. L'étude s'appuie sur des études antérieures qui comparent des trigrammes dans des textes anglais originaux de fiction avec des textes de fiction traduits du norvégien. À l'enquête actuelle s'ajoutent des textes anglais traduits de deux langues supplémentaires – l'allemand et le suédois – dans le but d'établir dans quelle mesure les tendances constatées dans les textes anglais traduits du norvégien sont comparables aux tendances observées dans des textes traduits d'autres langues. Ainsi, l'étude contribue à la discussion des universaux de la traduction et de la traduction comme troisième code. Sur le plan des fonctions de trigrammes, on a découvert que les originaux et les traductions anglaises partagent des caractéristiques fonctionnelles similaires dans huit des quatorze catégories identifiées. Sur les six autres, quatre montrent des différences statistiquement significatives entre les originaux et les traductions, quelle que soit la langue source. Une étude plus qualitative de quatre trigrammes spécifiques de deux de ces catégories conclut, comme c'est le cas pour les études précédentes, que ceci probablement s'explique par le fait que des éléments de la langue source transparaissent dans la traduction, et que les traducteurs ont une tendance (potentiellement universelle) à utiliser un ensemble relativement limité et fixe d'expressions dans leurs traductions.

The function of recurrent word-combinations in English translations from three different languages

SIGNE OKSEFJELL EBELING
University of Oslo, Oslo, Norway
s.o.ebeling@ilos.uio.no

RÉSUMÉ

Cet article compare les tendances phraséologiques dans les textes anglais traduits et non traduits en utilisant des séquences de trois mots classés selon leur fonctionnement. L'étude s'appuie sur des études antérieures qui comparent des trigrammes dans des textes anglais originaux de fiction avec des textes de fiction traduits du norvégien. À l'enquête actuelle s'ajoutent des textes anglais traduits de deux langues supplémentaires – l'allemand et le suédois – dans le but d'établir dans quelle mesure les tendances constatées dans les textes anglais traduits du norvégien sont comparables aux tendances observées dans des textes traduits d'autres langues. Ainsi, l'étude contribue à la discussion des universaux de la traduction et de la traduction comme troisième code. Sur le plan des fonctions de trigrammes, on a découvert que les originaux et les traductions anglaises partagent des caractéristiques fonctionnelles similaires dans huit des quatorze catégories identifiées. Sur les six autres, quatre montrent des différences statistiquement significatives entre les originaux et les traductions, quelle que soit la langue source. Une étude plus qualitative de quatre trigrammes spécifiques de deux de ces catégories conclut, comme c'est le cas pour les études précédentes, que ceci probablement s'explique par le fait que des éléments de la langue source transparaissent dans la traduction, et que les traducteurs ont une tendance (potentiellement universelle) à utiliser un ensemble relativement limité et fixe d'expressions dans leurs traductions.

ABSTRACT

This article compares phraseological tendencies in translated vs. non-translated English through functionally classified 3-word sequences. The study builds on previous research that compared 3-grams in fiction texts originally written in English with fiction texts translated from Norwegian. The current investigation adds English translations from two additional languages – German and Swedish – with the aim of establishing to what extent the tendencies noted for English translations from Norwegian extend to English translations from other languages. Thus the study contributes to the discussion of translation universals and translation as a third code. At the level of 3-gram functions, it has been uncovered that English originals and translations share similar functional characteristics in eight of the fourteen categories identified. Of the remaining six, four show statistically significant differences between originals and translations, regardless of source language. Based on a more qualitative study of four specific 3-grams from two of these categories, it is concluded, in line with the previous studies, that the most likely explanations are source language(s) shining through and the (potentially universal) tendency for translators to use a smaller and more fixed set of expressions in their translations.

RESUMEN

Este artículo compara las tendencias fraseológicas en inglés traducido frente a inglés no traducido a partir de secuencias de 3 palabras clasificadas funcionalmente. El estudio se basa en investigaciones previas que comparan trigramas en textos de ficción escritos originalmente en inglés con textos de ficción traducidos al inglés del noruego. La investigación actual agrega traducciones al inglés de dos idiomas adicionales, alemán y sueco, con el objetivo de establecer en qué medida las tendencias observadas para las traducciones al inglés del noruego se extienden a las traducciones al inglés de otros idiomas. Este estudio contribuye así a la discusión de los universales de la traducción y la traducción como un tercer código. A nivel de funciones de trigramas, se ha descubierto que los originales y las traducciones en inglés comparten características funcionales similares en ocho de las catorce categorías identificadas. De las seis restantes, cuatro muestran diferencias estadísticamente significativas entre los originales y las traducciones, independientemente del idioma de origen. Basado en un estudio más cualitativo de los cuatro trigramas específicos de dos de estas categorías, se concluye, en línea con los estudios previos, que las explicaciones más probables son el idioma o idiomas de origen que se hacen visibles y la tendencia (potencialmente universal) de los traductores a utilizar un conjunto de expresiones más pequeño y más fijo en sus traducciones.

MOTS-CLÉS/KEYWORDS/PALABRAS CLAVE

phraséologie, combinaisons de mots récurrentes, classification fonctionnelle, anglais traduit vs non traduit, différentes langues sources, fiction

phraseology, recurrent word-combinations, functional classification, translated vs. non-translated English, different source languages, fiction

fraseología, combinaciones de palabras recurrentes, clasificación funcional, inglés traducido frente a no traducido, diferentes idiomas de origen, ficción

1. Introduction

This study investigates phraseological tendencies in English original (EO) and translated (ET) texts, with the aim of shedding light on the use of functionally defined sequences of words in EO vs. ET. In two previous studies of recurrent word-combinations in EO and ET, no significant differences were observed for more than half of the functional categories identified (Ebeling and Ebeling 2017, 2018). In a more detailed study of two of the categories that were found to differ (Comparison, e.g. *as good as* and Spatial, e.g. *across the table*), Ebeling and Ebeling (2017) tentatively attributed the differences to the source language shining through – in this case Norwegian. Drawing on these findings, and applying the same method, the current study classifies and analyses recurrent word-combinations in English translations from two additional source languages – German and Swedish – with the purpose of establishing with more certainty the extent to which it is the source language that influences the functional makeup of the translations or whether it can be attributed to more general characteristics of translated language, regardless of source language.

The primary data for the previous studies were compiled from the English-Norwegian Parallel Corpus English+ and will serve as a basis for the comparison with the new material from the English-Swedish Parallel Corpus (ESPC) and the Oslo Multilingual Corpus (OMC), both of which contain English original and translated fiction texts (from Swedish and German sources, respectively). The choice of source languages from which the English translations originate is to some extent a pragmatic one. In order to make a comparison with the studies referred to above, comparable corpus data from fiction are needed and such data are available in the ESPC and OMC. Moreover, some knowledge of all the languages involved, on the part of the researcher, is required to facilitate the task of interpreting the results.

The method applied can be said to be "knowledge-free" (Baroni and Bernardini 2003: 85), in the sense that uninterrupted sequences of three words (3-grams) will be extracted automatically from the respective corpora. The functional classification of the 3-grams draws on Altenberg (1998), in particular, but is also inspired by Moon (1998) and Biber, Conrad, *et al.* (2004). The taxonomy operates with four main functional classes: Evaluative, Informational, Modalising, and Organisational, with Informational being further divided into 12 sub-categories. The functionally classified 3-grams will then undergo a quantitative comparison in a two-tailed *t*-test implemented in R[1] to establish to what extent the translations from different source languages are similar or (significantly) different from each other. As the source languages German, Norwegian and Swedish are relatively closely related Germanic languages, we may not expect translations from these into another closely related language – English – to significantly differ from each other at the level of 3-gram functions. However, as many researchers have pointed out, translated language seems to embody some characteristics that set it apart from non-translated language (for example Frawley 1984; Teubert 1996; Mauranen 1998; Baker 2004; Halverson 2017, to mention a few).

The study is firmly placed within a tradition of translation studies in which the comparison of translated vs. non-translated text in the same language is central, for instance Teich (2003) on German and English; Baker (2004, 2007) on English; Xiao (2011) on Chinese; Lee (2013) on Korean; De Baets, Vandevoorde, *et al.* (2020) on Dutch. It takes a phraseological approach and adds the dimension of combining corpus-linguistic methods with a general and systematic functional analysis and is as such concerned with broader categories rather than specific, predefined items. Findings from a study like the current one are important in the wider context of translation studies in that it contributes new insights into the function of word sequences in translated language and problematises the view that translated language necessarily constitutes a "third code" (Frawley 1984).

This article has the following structure: Section 2 gives an account of the material and method used. In Section 3, the previous studies are presented in more detail, including important observations and results, before moving on to the analysis of 3-grams in the English translations from German, Norwegian and Swedish (Section 4). Section 5 offers a comparison between the functions of 3-grams in EO and ET from the three different languages, as well as more detailed, qualitative studies of two of the categories that yield statistically significant results between all three translation pairs. Some concluding remarks are given in Section 6.

2. Material and method

As mentioned, this study applies corpus-linguistic methods to extract data from three different corpora and is thus in line with a framework that advocates "greater methodological rigour in corpus-based translation studies," viz. the Contrastive Translation Analysis (CTA) approach (Granger 2018: 189). I will start by outlining the CTA model in Section 2.1, before moving on to a description of the corpora in Section 2.2. The data extraction method is explained in Section 2.3, followed by a brief description of the functional classification procedure in Section 2.4.

2.1. Contrastive translation analysis

Granger (1996) has previously devised the Integrated Contrastive Model (ICM), which is a methodological framework that combines Contrastive Analysis and Contrastive Interlanguage Analysis. More recently, she has adapted the model to a Translation Studies perspective in the hope that it "could contribute to strengthening the empirical basis of the field" (Granger 2018: 189). This is clearly achieved by incorporating a wide range of corpora for the purpose of Contrastive Translation Analysis. Figure 1 is based on Granger's representation of the model and includes the corpus design with English as the focal language, but with the languages of the current study explicitly incorporated into the model, that is German, Norwegian and Swedish. Granger's CTA model is robust in the sense that it integrates English translation data with their respective source texts, as well as comparable English original texts. In addition, the model opens for a comparison with larger monolingual comparable corpora (of original texts) of all the languages involved.[2] Finally, texts written by learners of English, whose mother tongues correspond to the languages involved, are also included in Granger's framework.

Figure 1

Contrastive translation analysis (based on Granger 2018: 190)

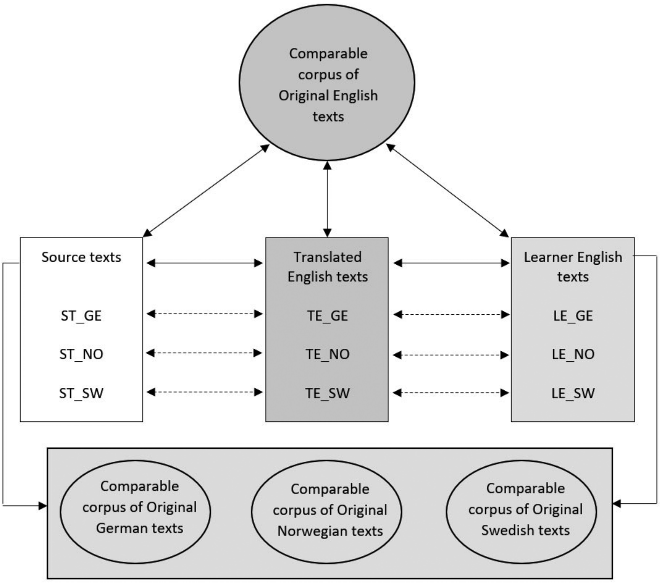

Comparable corpus of Original English texts
Source texts
ST_GE
ST_NO
ST_SW
Translated English texts
TE_GE
TE_NO
TE_SW
Learner English texts
LE_GE
LE_NO
LE_SW
Comparable corpus of Original German texts
Comparable corpus of Original Norwegian texts
Comparable corpus of Original Swedish texts

Figure 1 differs from Granger's original model in that the middle part has been highlighted (dark grey) and two parts have been shaded in grey. The former is the focus of the current study, where the functions of 3-grams in translations into English from different languages will be compared with each other and with texts originally written in English. The latter are parts that are not relevant to this investigation. The white box to the left, however, is relevant, as the German, Norwegian and Swedish source texts are available for the more qualitative observations in Sections 5.1 and 5.2. Moreover, Granger (2018: 189) notes that "[t]he full CTA model is bidirectional" and is as such reminiscent of both Granger's (1996) Contrastive Analysis section of the Integrated Contrastive Model and Johansson's bidirectional corpus structure developed for the English-Norwegian Parallel Corpus (ENPC) (Johansson and Hofland 1994). Although the ENPC is one of the sources of data in this study, the bidirectionality of the corpus (including translations from English into Norwegian) is not relevant here. Thus, it is primarily a unidirectional study, with the added dimension of comparing translated English with texts originally written in English.

2.2. The corpora

In the previous studies where recurrent word-combinations in texts originally written in English were compared with texts translated into English from Norwegian (Ebeling and Ebeling 2017, 2018), the English-Norwegian Parallel Corpus English+ (ENPC+) was used as the primary source of data. As the same corpus will be used in the current investigation with a more extensive set of 3-gram types, a brief description of the corpus is in order. The ENPC+ is an expanded version of the fiction part of the English-Norwegian Parallel Corpus (Johansson and Hofland 1994). It is a bidirectional corpus consisting of English and Norwegian original texts and their translations into Norwegian and English. Of the 39 original texts in each language, 30 are identical to the 10,000-15,000-word text extracts constituting the ENPC, while the remaining nine texts in Norwegian and eight of the English ones are full-length novels. The 39th English text is also a shorter text extract of approx. 12,000 words. The sub-corpora of the ENPC+ relevant to the current study – English originals and English translations – contain, in total, approx. 1.3 and 1.4 million words, respectively. The texts were published in the period from 1980 to 2012. For a fuller description of the ENPC, see Johansson, Ebeling, *et al.* (1999/2001),[3] and of the ENPC+, see Ebeling and Ebeling (2013).[4]

Furthermore, corpus material is culled from two other sources: the English-Swedish Parallel Corpus (ESPC) for English translations from Swedish and the Oslo Multilingual Corpus (OMC) for English translations from German. The ESPC is the sister corpus of the ENPC and the two corpora have the same bidirectional structure. The fiction part of the ESPC contains 25 English and Swedish original texts extracts of 10,000-15,000 words and their translations. With a few exceptions, all the texts were published in the 1980s and 1990s. For the purpose of this study, it is mainly the English translations that will be analysed, but the Swedish originals will also be consulted. In total, these two sub-corpora consist of roughly 330,000 words (translations from Swedish) and 310,000 words (Swedish originals). For a more detailed description of the ESPC, see Altenberg and Aijmer (2000).[5]

The Oslo Multilingual Corpus (OMC) "is a collection of text corpora comprising original texts and translations from several languages," including Dutch, English, French, German, Norwegian and Portuguese.[6] The sub-corpora constituting the OMC are either bi-, tri- or unidirectional and the English-German sub-corpus used here is of the bidirectional type, with originals and translations in both directions. The full English-German sub-corpus contains 33 English and 21 German original fiction and non-fiction text extracts of roughly 15,000 words each, mainly published in the 1990s. However, the restriction of including only fiction texts meant that the current study is based on a much smaller portion of this sub-corpus, namely

10 German originals with their English translations, amounting to a total of around 150,000 words in each section. The English-German sub-corpus is referred to as "En-Ge-En" within the OMC family of corpora but will be referred to here as the OMC. See Johansson (2002) for a more detailed description of the OMC.[7]

An important feature of all the corpora used is that the texts have been written/translated by a number of different authors/translators, thus avoiding a strong idiosyncratic bias in favour of one particular author/translator. In other words, the corpora are sampled to include a variety of different informants.

Table 1 gives an overview of the number of texts and running words in the most relevant sub-corpora, i.e. English translations from German (ET_GE), English translations from Norwegian (ET_NO), English translations from Swedish (ET_SW),[8] and English originals (EO). As the quite marked differences in size between the corpora may have an impact on the results of the study, it is important to keep this in mind in the later analysis and interpretation of the data.

Table 1

Comparison of corpora used in terms of number of texts and running words

	OMC	ENPC+	ESPC	ENPC+
	ET_GE	ET_NO	ET_SW	EO
Number of texts	10	39	25	39
Number of words	154,025	1,418,321	337,637	1,316,006

With reference to the final paragraph in Section 2.1, it is worth reiterating that, although all three corpora are bidirectional, this investigation draws only on unidirectional material and is thus focused on the dark grey and white squares of Figure 1. Moreover, comparisons with 3-grams from texts originally written in English are based on the English original texts in the ENPC+, as indicated in Table 1.

2.3. Data extraction

The studies upon which the present one is based (Ebeling and Ebeling 2017, 2018) investigated uninterrupted sequences of three words (3-grams). This was in many ways an experimental approach to see how the function of such sequences could be compared in original vs. translated texts, thus going beyond a traditional comparison of individual lexemes. Such an approach is also in line with the view that language to a large extent is phraseological in nature, that is language may be said to rely on "combinations of words that customarily co-occur" (Kjellmer 1991: 112). The choice of sequences of exactly three words as the object of study was motivated primarily by the fact that the corpus used was relatively small and would produce few recurrent 4- or 5-grams. As the size of the corpora is even more of a challenge in the current study, 3-grams continue to be the focus in the comparison between English originals and translations from German, Norwegian and Swedish. This is in agreement with Culpeper and Kytö's claim that 3-grams "offer a good compromise between the great number of different two-word combinations and the small number of different four-word combinations" (Culpeper and Kytö 2002: 45).

To ensure that the material is representative of the several authors and translators in the corpus, quite rigid extraction thresholds of dispersion and recurrence were set. First, a 3-gram type must occur in at least twenty-five per cent of the texts, i.e.in other words in ten out of the thirty-nine texts in the English originals and the translations into English from Norwegian (ET_NO). 3-grams in translations from German are required to occur in at least three different texts and, in translations from Swedish, in at least six different texts. Moreover, an additional, quite conservative threshold requiring each 3-gram to occur with a frequency of at least twenty per million words (pmw) was implemented in the original studies (that is, twenty-six and twenty-eight times in each of the sub-corpora: EO and ET_NO, respectively).[9] For translations from Swedish, this means a minimum of seven occurrences and, from German, a minimum of slightly more than three, adjusted to four for the purpose of this study. It should be noted that "even if a 3-gram was not frequent enough or did not occur in at least twenty-five per cent of the texts in one of the sub-corpora, this does not mean that it was not attested at all in that sub-corpus" (Ebeling and Ebeling 2018: 352).

The computer software AntConc (Anthony 2019)[10] is used to extract 3-gram types in the material; this was also the case in the previous studies. In addition to the above-mentioned thresholds, some changes to the default settings in AntConc were made to ensure that: (1) tags/mark-up were not part of the 3-grams; (2) apostrophe and hyphen were not treated as word delimiters; and (3) 3-grams did not run across s-unit (sentence) boundaries.

This method of data extraction produces comparable lists of 3-gram types from the different corpora, one for the EO texts and three for the translated texts, amounting to 1,408 (EO), 1,371 (ET_GE), 1,468 (ET_NO) and 1,149 (ET_SW) 3-gram types, respectively (see Section 4). When scrutinising the lists, it soon became evident that the 3-gram types that made the threshold in the EO texts, but not in the ET texts, were in fact also attested in the translations, albeit not beyond the threshold. Similarly, the types that reached the threshold in the ET corpora were also attested further down the list in the other corpora. Thus, as was the case in the original studies (Ebeling and Ebeling 2017, 2018), the token counts are further evened out on the basis of 3-grams meeting the thresholds in one of the corpora. Table 2 illustrates this procedure on the basis of three 3-gram types, one of which reaches the threshold in all the corpora (*for a while*), one which reaches the threshold in two corpora (*for so long*) and one which only reaches the threshold in one corpus (*for many years*).

TABLE 2

Example of 3-gram types that were added after the initial extraction stage (raw number of occurrences and range in terms of number of texts)

3-gram	EO raw | range	ET_GE raw | range	ET_NO raw | range	ET_SW raw | range
For a while	152 | 25	21 | 9	188 | 31	32 | 18
For many years	5 | 4	1 | 1	34 | 13	6 | 3
For so long	21 | 11	4 | 3	30 | 14	5 | 4

The first number in the shaded cells in Table 2 indicates the number of tokens that were added for 3-gram types that initially did not reach the threshold in their respective corpora, either due to a low recurrence rate or to a narrow dispersion range (the second number in the cells in Table 2). The reasoning behind this was that by adding the types from the combined lists we would ensure a comparison of the same number of 3-gram types (Ebeling and Ebeling 2017: 40-41). The number of 3-gram types used in the present study is therefore 2,765.[11]

Following this extraction procedure, the 3-gram types are then classified into the functional categories that form the basis for the comparison between the texts originally written in English and the texts translated into English from three different languages. The token counts for each category provide the input used in the statistical tests (that is, the number of times a 3-gram type belonging to that category is encountered).[12]

2.4. *Functional classification of the 3-grams*

As pointed out in the Introduction, the functional classification of the 3-grams is primarily inspired by Altenberg (1998), but also by the taxonomies presented in Moon (1998) and Biber, Conrad, *et al.* (2004). In many ways the classification scheme adopted here can be seen as a fusion of elements from all three, with some adjustments to accommodate the current dataset. The four main functional classes are Evaluative, Informational, Modalising, and Organisational. Moreover, the Informational class is divided into 12 sub-categories. Table 3, a slightly adapted version of the one originally published in Ebeling and Ebeling (2018), gives an overview of all 15 categories, including a brief definition and examples of each. The 12 categories not highlighted in bold are all Informational, while the ones in bold represent the other three main categories.

Table 3
Functional categories of 3-grams in the material

Functional Category	Definition	Examples
Comparison	Expresses some kind of comparison	*as good as, as if to, looked like a*
Contingency	Expresses a condition, reason, cause or concession	*because it was, if he ‘d, why did you*
Evaluative	Similar to modalising but typically contains an evaluative adjective or adverb instead of a verb	*‘s a good, i ‘m sure, just do n’t*
Existential	Contains existential *there*	*and there ‘s, there were no*
Fragment	Typically consists of noun phrase(s) (fragments) that could be either thematic or rhematic. Some verb phrase(s) (fragments) are also found in this category	*a sense of, the door and, to go on*
Modalising	Contains verbs that are either identifiable as modal auxiliaries or other items (typically a verb) expressing attitude, possibility/probability or certainty towards a proposition	*‘ll tell you, but he could, seemed to be*
Organisational	Contains items that are clearly recognizable as text structuring devices	*all the same, in any case*
Process	Is represented by manner and means expressions	*in a way, the way you*
Quantifying / Intensifying	Contains quantifying and intensifying expressions	*a glass of, more or less, lot of time*
Reporting	Includes a reporting verb	*he said and, no he said*

Functional Category	Definition	Examples
Respect[13]	Includes abstract circumstances of the action identifying "a relevant point of reference in respect of which the clause concerned derives its truth value" (Quirk et al. 1985: 484)	*apart from the*
Rhematic	Typically includes a verb followed by (part of) a noun phrase (that is the beginning of an object or complement/predicative)	*'s not a, he told me, to give him*
Spatial	Includes a clear spatial reference	*across the table, back in the, to be there*
Temporal	Includes a clear temporal reference	*a few days, at the moment, he 'd never*
Thematic Stem	"Consist[s] of subject and verb (plus any preceding thematic elements) but lack[s] a rhematic post-verbal element" (Altenberg 1998: 111)	*and I'm, but he had, what's happened*

It is important to note that it is not always straightforward to classify 3-grams functionally, simply because three words may not be enough to establish what the function is. In cases of doubt, concordance lines (in essence the extended context of the sequences) were checked. Furthermore, it was decided that (potentially) ambiguous 3-grams could only belong to one category. Again, concordance lines were checked, this time to determine the most frequent use of the 3-gram in the material. A case in point is the following example from Ebeling and Ebeling (2017): the 3-gram *he was about* can potentially be analysed as Temporal, as in Example (1), or Spatial, as in Example (2). As the most frequent use attested in the corpus refers to time, *he was about* was classified as Temporal (Ebeling and Ebeling 2017: 18).[14]

1) He remembered the matches just as *he was about* to dive in and left them on the tiles. (ENPC+/MoAl1E[15])

2) *He was about* one step from patting my hand and calling me "love." (ENPC+/TaFr1E)

3. Previous studies (and steps in the current analysis)

The two previous studies of the functions of 3-grams in English originals vs. English translations (from Norwegian) referred to above (Ebeling and Ebeling 2017, 2018) draw on material from the ENPC+ and present a methodological framework, as well as results, relevant to the present study. Although one of the aims of both articles was to contribute to the discussion of the use of translation data in contrastive studies between languages, they have slightly different foci, notably more focus on the linguistic side of things in the former and on methodological issues in the latter.

The quantitative comparison of the functional categories in English originals vs. English translations (from Norwegian) presented in Table 4 overlaps in the two articles. An independent, two-tailed *t*-test with Welch's correction in R was used to compare these frequencies of 3-gram functions in EO vs. ET,[16] the results of which are shown in Table 4; significant *p*-values are highlighted in bold.[17]

Table 4
***P*-values calculated for each functional category (from Ebeling and Ebeling 2017)**

Category	***P*-value**	**Favoured in**
Comparison	**$p = 0.002$**	ET
Contingency	$p = 0.679$	--
Evaluative	$p = 0.210$	--
Existential	$p = 0.605$	--
Fragment	**$p = 0.001$**	ET
Modalising	$p = 0.313$	--
Organisational	**$p = 0.002$**	ET
Process	**$p = 0.016$**	ET
Quantifying/Intensifying	$p = 0.299$	--
Reporting	**$p = 0.005$**	EO
Rhematic	$p = 0.522$	--
Spatial	**$p < 0.001$**	ET
Temporal	**$p < 0.001$**	ET
Thematic Stem	$p = 0.991$	--

As can be seen in Table 4, seven of the 14 categories show a statistically non-significant result, which suggests that EO and ET behave similarly at this functional level of analysis for the categories Contingency, Evaluative, Existential, Modalising, Quantifying/Intensifying, Rhematic, and Thematic Stem. This was taken as a signal that the use of translations should not, as a matter of course, be regarded as a "third code" (cf. Ebeling and Ebeling 2017: 44).

As the other half did produce statistically significant results, a more qualitative, albeit non-exhaustive, investigation into some of those categories was carried out. For example, in the case of one Comparison 3-gram type (*it was as*) that contributed to the statistically significant difference between English original vs. English translated text, it was concluded that the huge difference in the token count for this 3-gram in EO vs. ET was a result of the translators' use of two 4-word sequences (*it was as if* and *it was as though*, the latter of which was not frequently attested in EO) and source language shining through, from the frequent Norwegian sequence det var som om that corresponds to precisely *it was as* (*if*) or *it was as* (*though*), boosting the number of occurrences for the 3-gram *it was as* in the translations (Ebeling and Ebeling 2017: 44). This suggests that translated language is different from non-translated language in some respects. However, it is important to stress all the 3-gram types that reached the threshold in EO were also attested in the ET and vice versa. Moreover, the studies included all 3-grams that reached the thresholds in either EO or ET, thus:

> [o]n the basis of the comparison of the token counts of the 3-grams extracted for the study, it seems that most differences are a matter of degree, rather than being systemic at the level of the functions investigated. (Ebeling and Ebeling 2018: 347)

The studies outlined in this section have laid the methodological foundation, and serve as the backdrop, for a similar quantitative analysis of 3-grams in English translations from three languages, in a direct comparison with the functional analysis of 3-grams in English originals. In this manner, the current study is an attempt at meeting the need to study a "variety of languages and language pairs" in order to discuss potential universal features of translation (Mauranen 2007: 45).

While Mauranen (2007: 45) proposes to include language pairs that are both typologically distant and close, this study will, as mentioned in the Introduction, only be concerned with the latter type of language pairs.

4. Analysis of 3-grams in English translations from German, Norwegian and Swedish

As pointed out by De Sutter, Goethals, *et al.* (2012: 142) "the study of translation cannot be performed successfully without acknowledging source structures, texts, or languages." Moreover, Lefer (2012) points to the fact that the source language as a crucial factor in translation studies has now been re-established after a period where many scholars were concerned with features that typically occur in translated texts and "which are not the result of interference from specific linguistic systems" (Baker 1993: 243). In accordance with these trends, the ETs have been split into different datasets according to source language and, to complement this, some specific source sequences have been selected for discussion in Sections 5.1 and 5.2.

In Section 2.3 the number of 3-gram types that meet the thresholds in the different sub-corpora was established. However, following the approach taken in the previous studies of English originals vs. translations from Norwegian, it is the combined number of 3-gram types reaching the thresholds in each of the corpora that form the basis for the further analysis. This also means that the 3-gram lists from the original studies were topped up with additional 3-gram types that met the threshold in the English translations from German and Swedish (see Section 2.3). Thus, the current study includes an additional 854 3-gram types compared to the original studies (2,765 types vs. 1,911). Table 5 shows the number of 3-gram types that meet the thresholds within the corpora, the number of 3-gram types after topping up, followed by the number of 3-gram tokens produced by the 2,765 types in each corpus. Finally, the total number of 3-gram tokens in each corpus is found in the right-most column to get a sense of the proportion of high-frequency 3-grams in the sub-corpora.

Table 5
3-gram types and tokens in the sub-corpora

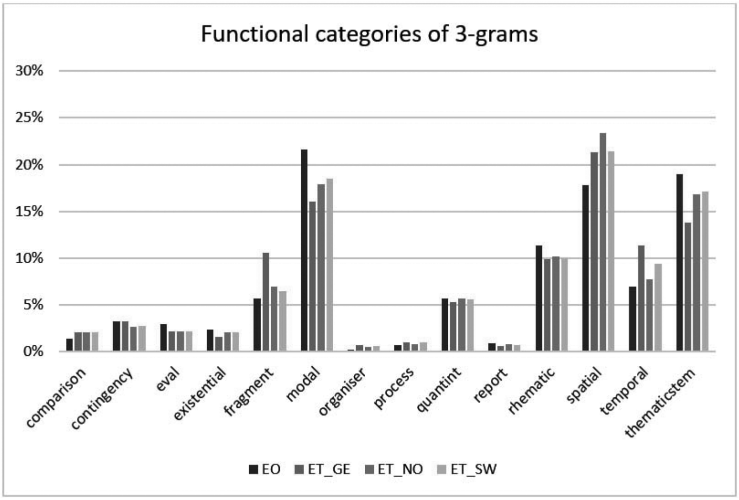
	# of 3-gram types (threshold types)	# of 3-gram types after top-up	# 3-gram tokens	Total # of 3-gram tokens in sub-corpus
EO	1,408	2,765	102,044	1,110,300
ET_GE	1,146	2,765	10,693	136,720
ET_NO	1,468	2,765	108,545	1,119,699
ET_SW	1,448	2,765	27,683	288,490

Despite the differences in corpus size, and therefore also different absolute thresholds for extraction (see Section 2.3), the sub-corpora behave similarly with regard to token proportions: the number of 3-gram tokens account for 7.8% (ET_GE), 9.7% (ET_NO) and 9.5% (ET_SW) of the total number of 3-gram tokens in the corpora. Admittedly, these are relatively modest proportions, but they do reflect the most frequently occurring and most evenly dispersed sequences in the corpora. A functional analysis of these will therefore be welcome in order to give an indication of the overall functional makeup of the texts in English original and translated texts.[18]

Figure 2 shows the percentagewise distribution of the 3-gram tokens according to function in the four corpora, comparing the distribution in English originals and translations into English from the three languages.

Figure 2

Distribution of 3-grams according to functional category (%)



What can be gleaned from this initial and crude overview of the functional distribution in the four datasets is that there is a relatively even proportional distribution in most of the functional categories, but with some notable exceptions. In terms of percentages, Fragment and Temporal 3-grams are markedly more frequent in translations from German, Modalising and Thematic Stem 3-grams are more frequent in English original texts and Spatial 3-grams are more frequently attested in translations overall compared to original texts. In order to get a more statistically sound picture of these differences, token counts for each functional category per text were counted and normalized. These counts form the basis for all the statistical tests performed in this study. Ebeling and Ebeling (2017) illustrate the procedure of normalising the counts with reference to two functional categories (Fragment and Modalising) in four texts (MoAl1E, MW1E, JoNe1TE, JW1TE) in the ENPC+, and is repeated here as Table 6.

Table 6

Token counts and normalized frequencies (Ebeling and Ebeling 2017: 40)

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

The normalised frequencies in Table 6 show the number of 3-gram tokens that meet the threshold per 1,000 3-gram tokens in a text. For instance, the (raw) number of Modalising tokens in the text MW1E is 158. This number is divided by 8,878 (= # of 3-gram tokens), resulting in a normalized frequency of 17.8 per 1,000 3-grams.

On the basis of such counts in the respective sub-corpora, the following sections present results from pairwise statistical tests for each of the functional categories between English original and translations from German (4.1), Norwegian (4.2) and Swedish (4.3), respectively.

4.1. English translations from German

From Figure 2 we can hypothesise that there may be some statistically significant differences in the use of some of the functional classes of 3-grams between English originals and translations from German. In particular, this may be expected for Fragment, Modalising, Spatial, Temporal, and Thematic Stem, but also some of the other categories may yield statistically significant results.

As described in Section 3, the next step was to run a two-tailed (independent) *t*-test to compare the use of functional categories (based on tokens; cf. Table 6) in English originals vs. translations from German.[19] The *p*-values obtained either reject (significant *p*-value) or fail to reject (non-significant *p*-value) the hypothesis that EO and ET_GE use functional categories of 3-grams with a similar frequency. Table 7 gives an overview of the results for each of the categories, including information on whether a category showing a significant *p*-value (**bold**) is favoured in the original or the translated texts (right-most column).[20]

Table 7

***P*-values calculated for each category: EO vs. ET_GE**

Category	***P*-value**	**Effect Size (Cohen's *d*)**	**Favoured in**
Comparison	*p* = 0.062	-0.842 (large)	
Contingency	*p* = 0.678	-0.163 (negligible)	
Evaluative	*p* = 0.085	0.437 (small)	
Existential	***p* = 0.012**	1.076 (large)	EO
Fragment	***p* < 0.001**	-2.445 (large)	ET_GE
Modalising	*p* = 0.168	0.567 (medium)	
Organisational	***p* = 0.012**	-1.569 (large)	ET_GE
Process	***p* = 0.012**	-1.426 (large)	ET_GE
Quantifying/Intensifying	*p* = 0.174	0.541 (medium)	
Reporting	***p* = 0.037**	0.636 (medium)	EO
Rhematic	*p* = 0.115	0.574 (medium)	
Spatial	*p* = 0.816	-0.088 (negligible)	
Temporal	***p* = 0.001**	-2.023 (large)	ET_GE
Thematic Stem	*p* = 0.076	0.673 (medium)	

Of the 14 categories, six are shown to differ significantly (with a medium to large effect size). As hypothesised on the basis of Figure 2, the Fragment and Temporal categories are significantly more frequent in English translations from German, whereas Spatial 3-grams are not, nor are Modalising 3-grams used significantly more in original texts compared to translations from German.

4.2. *English translations from Norwegian*

Turning now to a comparison between functional categories in English originals vs. translations from Norwegian, it will be interesting to see how the results relate to the results from the previous section, as well as the previous studies referred to above.

From Figure 2 and the previous studies of EO vs. ET_NO (Ebeling and Ebeling 2017, 2018), it may be expected that EO and ET_NO will differ significantly in their use of several functional categories, including Comparison, Fragment, Modalising, Organisational, and Spatial. Table 8 shows the results based on the same corpora as the previous studies, but, as described above, with a larger set of 3-gram types as input. In cases where the data were not normally distributed in either EO or ET_NO or both,[21] two statistical tests were run on the material, viz. *t*-test and Wilcoxon. In Table 8, only the *p*-values from the *t*-test are reported, since both tests showed similar tendencies.[22]

Table 8

***P*-values calculated for each category: EO vs. ET_NO**

Category	***P*-value**	**Effect Size (Cohen's *d*)**	**Favoured in**
Comparison	**$p < 0.001$**	-0.796 (medium)	ET_NO
Contingency	$p = 0.936$	0.018 (negligible)	
Evaluative	$p = 0.346$	0.215 (small)	
Existential	$p = 0.548$	0.137 (negligible)	
Fragment	**$p < 0.001$**	-0.781 (medium)	ET_NO
Modalising	$p = 0.437$	0.177 (negligible)	
Organisational	**$p = 0.007$**	-0.630 (medium)	ET_NO
Process	**$p = 0.006$**	-0.640 (medium)	ET_NO
Quantifying/Intensifying	$p = 0.165$	-0.318 (small)	
Reporting	$p = 0.301$	0.236 (small)	
Rhematic	$p = 0.561$	0.132 (negligible)	
Spatial	**$p < 0.001$**	-0.888 (large)	ET_NO
Temporal	**$p < 0.001$**	-0.999 (large)	ET_NO
Thematic Stem	$p = 0.802$	-0.057 (negligible)	

Not unexpectedly, the results shown in Table 8 are very much in line with the previous studies regarding categories that differ significantly between EO and ET_NO (see Table 4). The only exception is Reporting, which, with the added 3-gram types in the current study, produces a non-significant *p*-value, albeit with a small effect size. It is also interesting to note that the 3-grams in all six categories with a significant result show an overrepresentation in the translated texts.

4.3. *English translations from Swedish*

We now turn to the comparison between functional categories in English originals vs. translations from Swedish. It can be hypothesised from Figure 2 that EO and ET_SW will differ significantly in their use of the functional categories Fragment, Modalising, Report, Spatial, and Temporal. In the five categories where either EO or ET_SW or both were not normally distributed,[23] both the *t*-test and Wilcoxon were run with similar results. Table 9 shows the results of the *t*-test.

Table 9
***P*-values calculated for each category: EO vs. ET_SW**

Category	***P*-value**	**Effect Size (Cohen's *d*)**	**Favoured in**
Comparison	$\boldsymbol{p < 0.001}$	-1.054 (large)	ET_SW
Contingency	$p = 0.147$	-0.364 (small)	
Evaluative	$p = 0.915$	-0.025 (negligible)	
Existential	$p = 0.435$	-0.212 (small)	
Fragment	$\boldsymbol{p < 0.001}$	-0.972 (large)	ET_SW
Modalising	$p = 0.187$	-0.347 (small)	
Organisational	$\boldsymbol{p < 0.001}$	-1.151 (large)	ET_SW
Process	$\boldsymbol{p = 0.008}$	-0.861 (large)	ET_SW
Quantifying/Intensifying	$p = 0.068$	-0.473 (small)	
Reporting	$p = 0.509$	0.169 (negligible)	
Rhematic	$p = 0.322$	-0.248 (small)	
Spatial	$\boldsymbol{p < 0.001}$	-0.867 (large)	ET_SW
Temporal	$\boldsymbol{p < 0.001}$	-2.082 (large)	ET_SW
Thematic Stem	$p = 0.086$	-0.433 (small)	

Table 9 further shows that, of the categories that appear to differ the most in Figure 2, three are found to differ significantly between originals and translations from Swedish, viz. Fragment, Spatial and Temporal. In addition, significant differences can be observed in the categories Comparison, Organisational, and Process, all with a medium to large effect size. In each of the categories showing a statistically significant result, the respective 3-grams are overrepresented in translated texts compared to originals, as can also be gleaned from the bar chart in Figure 2.

5. Comparison of 3-gram functions in English original texts vs. English translated texts from three different source languages

The previous sections outlined the functions of 3-grams in texts translated into English and texts originally written in English. In all three original vs. translation pairs, six categories were found to differ significantly between translated and non-translated English, while eight categories were not. Table 10 juxtaposes the pairwise comparisons, with indications of significance. It is also worth noticing that, with the exception of Organisational, it is the Informational sub-categories (non-bold: Existential, Fragment, Process, Report, Spatial, Temporal) that represent these differences.

Table 10

Statistically significant (*) vs. non-significant (-) results for the 14 categories in translations from three languages into English

Category	EO vs ET_GE	EO vs ET_NO	EO vs ET_SW
Comparison	-	***	***
Contingency	-	-	-
Evaluative	-	-	-
Existential	*	-	-
Fragment	***	***	***
Modalising	-	-	-
Organisational	*	**	***
Process	*	**	**
Quantifier/Intensifier	-	-	-
Report	*	-	-
Rhematic	-	-	-
Spatial	-	***	***
Temporal	***	***	***
Thematic Stem	-	-	-

Table 10 reveals both similarities and differences between the three EO-ET pairs. The main focus will be on the similarities, since space does not allow me to go into the differences in much detail. However, there are two categories that return a statistically significant result that are unique to German: Existential and Report. Both are more frequently attested in original English texts than in translations from German and it could be inferred that German uses fewer existential constructions than English, thus giving rise to fewer existential 3-grams in the translations. To speculate about the reason for the overrepresentation of Report 3-grams in English originals, it could be that the English texts contain more dialogue, and thus more reporting clauses will emerge in the originals. These are mere speculations and will have to await further study. Table 10 also shows that there are two categories – Comparison and Spatial – that show a significant difference in translations from Norwegian and Swedish, but not in translations from German. This is in line with Ebeling and Ebeling's (2017: 31) observation that "the more frequent use of Comparison and Spatial 3-grams in ET [from Norwegian] is most likely a result of source language shining through."

Regarding similarities across the corpora, it is interesting to note that there are six categories that do not produce significant differences in any of the EO-ET pairs: Contingency, Evaluative, Modalising, Quantifier/Intensifier, Rhematic, and Thematic Stem. This suggests that these are relatively stable functions in English fiction, at least in a Western context, as the categories do not discriminate between originals and translations at this level of expression. In other words, these functions are preserved naturally in translations from the three languages represented here. However, this is not the case in the four categories boasting a statistically significant difference in all three EO-ET pairs, namely Fragment, Organisational, Process, and Temporal. The similarity between the EO-ET pairs lies in the more frequent use of these four categories in the translated texts compared to their use in the original texts. There may be several reasons for this and it is particularly tempting to speculate that, at least in the case of Organisational and Temporal 3-grams, there may be an explicitation effect, whereby the translators more explicitly refer to the organisation of the text and give temporal cues.

As the scope of this paper does not allow an in-depth investigation of all four categories, I will restrict the following analysis to a more detailed examination of the categories Organisational (Section 5.1) and Process (Section 5.2). Focusing on the most frequently occurring 3-grams in the translated texts, I will search for specific sequences in the translations in the different corpora to establish what the source texts can tell us about potential reasons for the frequent use of these categories in translations into English.[24] It is tempting to suggest that the three source languages have something in common triggering these functions, in other words, this set of source languages seem to be subject to the same "gravitational pull" (Halverson 2017),[25] resulting in what may be perceived as explicitation in the case of Organisational 3-grams.

5.1. Analysis of two Organisational 3-gram types

There are 12 Organisational 3-gram types in the material. These are listed in Table 11 according to their raw token frequency in English originals and the top five in each of the corpora are shaded in grey.[26] Of these, two are found among the top five in translations from all three languages as well as in the original English texts: *the other hand* and *all the same.*

Table 11

Organisational 3-grams: types and tokens (raw frequency)

3-gram types	**EO (tokens)**	**ET_GE (tokens)**	**ET_NO (tokens)**	**ET_SW (tokens)**
by the way	47	3	77	17
in any case	40	11	45	22
the other hand	38	13	97	30
at any rate	30	4	89	6
all the same	30	18	68	30
in other words	16	7	49	11
in that case	12	4	29	13
all in all	12	4	9	2
as it were	8	4	12	13
on the contrary	7	7	47	11
other hand he	4	1	10	8
and in fact	3	5	6	0
Total (raw and per 100,000 3-grams)	247 (22.2)	81 (59.2)	538 (48.1)	163 (56.5)

A closer look at *the other hand* and *all the same* reveals that, in the great majority of cases, the 3-grams have an overt correspondence in the source texts. Not unexpectedly, *the other hand* is most commonly part of the 4-gram *on the other hand*, which is a preferred translation of German **andererseits** (Example 3), Norwegian derimot or på den annen side (Example 4) and Swedish *å andra sidan* or *däremot* (Example 5).

3) *On the other hand* I've always loved her, with all her dreadful faults – loved her and hated her.
(OMC/TBE1TE)[27]

Andererseits, ich habe sie immer geliebt, mit allen ihren Fürchterlichkeiten.
(OMC/TBE1)

4) *On the other hand*, I am not so far from Loaf's headquarters in Grønlandsleiret.
(ENPC+/PeRy1TE)

På den annen side er jeg ikke så langt fra Loffs hovedkvarter i Grønlandsleiret.
(ENPC+/PeRy1N)

5) *On the other hand*, he seemed to feel free to roam around the room.
(ESPC/LH1T)

Däremot kände han sig tydligen inte förhindrad att fritt röra sig i rummet.
(ESPC/LH)

Even if Examples (3)-(5) show the main trend in terms of correspondences, there are a couple of further observations worth making. First, both the Norwegian and Swedish sources trigger "false" hits in the translations, in the sense that not all instances of *the other hand* have an Organisational function. Rather, the literal sense of *hand* occurs in two out of the 30 cases in the translations from Swedish and in six out of 97 cases from Norwegian, as shown in Example (6).

6) *The other hand* – his right – was holding the remote control.
(ENPC+/JoNe1TE)

Den andre hånden – høyre hånd – holdt rundt en fjernkontroll.
(ENPC+/JoNe1N)

More interestingly, perhaps, there are a few cases of zero correspondences (in other words there is no overt source), six in translations from Norwegian, Example (7) and three from Swedish, Example (8).

7) On the other hand, his distrust of human beings appeared to grow…
(ENPC+/KH1T)

Hans frykt for menneskene ser ut til å vokse.
[Ø His fear of human beings seems to grow]
(ENPC+/KH1)

8) Palloo, on the other hand, regarded time as his most reliable accomplice.
(ESPC/LH1T)

Palloo hade tiden som sin trognaste bundsförvant.
[Palloo Ø had the time as his most faithful ally]
(ESPC/LH1)

Similar patterns of correspondence emerge for the 3-gram *all the same*. In this case, Norwegian and Swedish have one clearly preferred source, namely likevel and *ändå*, respectively, – both in an organisational concession sense – and each accounting for more than half of the occurrences.[28] In the German sources, it is either **trotzdem** or (**aber**) **doch** that typically gives rise to *all the same*. Moreover, there are some instances of non-organisational *all the same* in the translations, particularly from German, as in (9). In fact, half of the instances (nine) in translations from German are of this kind, of which seven are from the same text.

9) "If it's really *all the same* to you," Atreyu argued, "you might just as well tell me." (OMC/ME1TE)

"Wenn es dir wirklich *ganz gleich* ist," drang Atréju in sie, "dann könntest du es mir ebensogut sagen."
(OMC/ME1)

Zero correspondences of *all the same* are rare in translations from all three languages, none in translations from Swedish, one from German, and two from Norwegian, one of which is shown in Example (10).

10) "Stop letting it bother you," Johanne thought, exhausted, but couldn't stop *all the same*.
(ENPC+/AnHo1TE)

Slutt å bry deg, tenkte Inger Johanne matt, men maktet det ikke.
[Stop meddling, thought Inger Johanne, faint, but wasn't able to Ø]
(ENPC+/AnHo1N)

This admittedly non-exhaustive analysis indicates that there is not enough evidence to suggest that the overrepresentation of Organisational 3-grams in English translations is a result of explicitation through the insertion of overt cohesive ties. For two of the most frequent 3-grams, it is quite clear that they, with a few exceptions, stem from overt items in the source texts, be they German, Norwegian or Swedish. This tentative conclusion does in some ways tie in with Fabricius-Hansen's (2005) findings with regard to the use of connectives in English, German and Norwegian. Although connectives and organisational items may not (always) refer to the same type of items, they definitely share some text organising functions. On the basis of data from the OMC, Fabricius-Hansen suggests the following "rules" for English and German, with Norwegian somewhere in between:

> English: If the informational effect of using the connective is rather low, then don't use it! ("Be brief!")
> German: If using the connective is more informative than not using it, then use it! ("Be precise!") (Fabricius-Hansen 2005: 43)

Similarly, in a number of studies, the GECCo[29] project team at Saarland University has uncovered contrasts between English and German when it comes to cohesive devices, including "cohesive conjunction," which are clearly related to, albeit not completely identical to, the Organisational category in this study. Nevertheless, Kunz, Degaetano-Ortlieb, *et al.* (2017: 293) find that German expresses conjunctive relations more often than English. Indeed, "a stronger tendency is attested for German towards explicitly expressing logico-semantic relations via conjunctive relations (particularly conjunctive adverbials) on the textual level" (Kunz, Degaetano-Ortlieb, *et al.* 2017: 303). Moreover, relating their findings to previously held assumptions of "a preference for more explicit strategies in German as compared to English," Kunz, Degaetano-Ortlieb, *et al.* (2017: 298) can corroborate this for the level of cohesion.

Based on the above discussion, it can be inferred that the overrepresentation of Organisational 3-grams in English translations may be attributed to a translation effect from all three languages. Such items are naturally more frequent in German, Norwegian and Swedish compared to English, thus, to some extent the source language(s) can be said to be shining through. However, it is also interesting to note that the translators seem to opt for a restricted number of (3-word) sequences in English, e.g. *all the same* and *the other hand*, as off-the-peg translations of a variety of sources. This is particularly the case for *all the same* in translations from Norwegian and Swedish. Apart from the most common sources mentioned (likevel and *ändå*), a number of other items are attested as sources of *all the same*, including *i alla fall* 'in all cases,' *men* 'but,' *trots allt* 'despite all' from Swedish and like fullt 'nevertheless' (lit.: [equally full]), for det 'even so' (lit.: [for that]), men 'but' from Norwegian. Such tendencies suggest, in accordance with previous studies, that frequent items occur more frequently in translated texts (Mauranen 2000: 10; Halverson 2017: 9; De Baets, Vandevoorde, *et al.* (2020).

5.2. Analysis of two Process 3-gram types

The Process category includes 3-grams that represent expressions of manner and means, and is one of the Informational sub-categories. In terms of types, this category is larger than the Organisational one and contains 37 different 3-gram types. While the Organisational category showed quite a bit of overlap between the corpora among the top five 3-grams, Process shows less overlap overall, as none of the top five 3-grams in any corpus is found among the top five in all. There are, however, pairwise overlaps that are illustrated in Table 12.[30]

Table 12

Top five Process 3-grams in the corpora

Rank	EO	ET_GE	ET_NO	ET_SW
1	**the way he**	by no means	*in a way*	*in a way*
2	the way she	the way she	so that the	***that's how***
3	*in a way*	in this way	***that's how***	the way it
4	the way I	in such a	in such a	**the way he**
5	the way you	and so on	***so that he***	***so that he***

Before we focus on the differences between the corpora, it is interesting to observe the homogeneous nature of the EO list. Four out of the top five are in effect the "same" 3-gram: *the way* + PRON. The same homogeneity is not noted for the translations and, as such, the ET lists do not represent the natural choice of 3-grams in this category.

The lack of overlap between the corpora regarding the most frequent 3-grams in this category made me opt for a different strategy here compared to the one adopted in Section 5.1. Instead of looking more closely at 3-grams found among the top five in all corpora, I searched for *the* most frequent 3-gram in translations from German (*by no means*), Norwegian and Swedish (*in a way*) in order to trace them back to their source items.

By no means occurs eight times in the translations from German, which relatively speaking makes it a lot more frequent than the eight occurrences in the English original texts (5.8 vs. 0.7 occurrences per 100,000 3-grams, respectively). The main source in the German originals is **keineswegs**, as in Example (11), accounting for five of the eight occurrences. The other source items, with one occurrence each, are **mitnichten**, **nicht so**, and **gar nicht**.

11) No, Oskar Alder was *by no means* a strict teacher.
(OMC/ROS1TE)

Nun war Oskar Alder *keineswegs* ein strenger Lehrer.
(OMC/ROS1)

In a way occurs 80 times (7.1 per 100,000 3-grams) in the translations from Norwegian and 22 (7.6 per 100,000 3-grams) in the translations from Swedish, compared to 51 times (4.6 per 100,000 3-grams) in English originals. There is some overlap regarding the Norwegian and Swedish sources. By far the most frequent source in Norwegian is på en måte, as shown in (12).

12) Or was crazy *in a way* that would defy standard definitions.
(ENPC+/JoNe1TE)

Eller var gal *på en måte* som ikke er allment akseptert.
(ENPC+/JoNe1N)

In Swedish, however, the most frequent source is *på sätt och vis* (13), while the Norwegian counterpart på sett og vis (alternatively på et vis) is the second-most frequent one.

13) *In a way* it made me happy.
(ESPC/RJ1T)

Det gladde mig *på sätt och vis*.
(ESPC/RJ1)

This correspondence begs the question of what the actual function of *in a way* is, as it, in the context presented in Example (13), seems to have more of a discourse function. This use is even more evident in Example (14) from Norwegian, where *in a way* is the translation of the discourse particle jo.

14) So I'm happy *in a way.*
(ENPC+/KaFo1TE)

Så jeg er *jo* glad.
[So I am |particle| happy]
(ENPC+/KaFo1N)

Clearly, then, *in a way* is a 3-gram that does not exclusively belong to the Process category and, although this may contribute to the overrepresentation of Process 3-grams in English translations, it also depends on the division of labour between the discourse use and Process use of *in a way* in original texts.[31]

There are also a couple of zero correspondences, as shown in Example (15).

15) But it was worse *in a way* because we couldn't see the knives.
(ESPC/RJ1T)

… fast här var det värre eftersom man inte såg knivarna.
[but here was it worse Ø because one didn't see the knives]
(ESPC/RJ1)

Example 15 suggests that the translator felt the need to hedge a quite direct statement in the Swedish original. In fact, Example 15 may be said to border on the discourse function of *in a way* as well.

The analysis of the two Process 3-grams is rather inconclusive and needs further investigation beyond what can be offered here. However, in terms of overrepresentation, the same tendency as for the Organisational 3-grams can be noted: source language shines through (as manner and means seem to be more frequently expressed in the source languages available here) and translators make use of a limited, but frequent, set of translations in their renderings of Process 3-grams. The latter observation is potentially a universal trend, although admittedly based on a limited set of translation pairs. In terms of more qualitative characteristics, on the other hand, the actual 3-grams that are most frequently used differ more between original and translated English in the Process category. This suggests that the differences in the use of Process 3-grams are both a matter of degree, since all 3-grams in the translations are attested in the original texts as well, but with a (much) lower recurrence, and to some extent quality, as shown in the overview of top five 3-grams in Table 12.

6. Conclusion

In Ebeling and Ebeling (2017: 47) one proposed avenue for further research was to incorporate "translated English fiction from other languages" than Norwegian, as this "would greatly enhance and extend the generality of studies of this kind." The current study has done exactly that, in including translations from German and Swedish. The analysis has to some degree enhanced and extended the generality of the results from the previous study.

First, findings from the previous studies were substantiated in the sense that (less than) half of the functional categories of 3-grams were found to differ significantly between texts originally written in English and texts translated into English from three different languages. This is an important finding in the context of translation studies, as it tells us that at this level of description (of functions of sequences of words rather than the use of individual lexical items), English translated and non-translated (fiction) texts behave in a very similar way. Second, of the six categories that were found to differ significantly, four were shared by all three translation pairs (EO-ET_GE, EO-ET_NO and EO-ET_SW), namely Fragment, Organisational, Process and Temporal. A detailed analysis of all these categories was not feasible within the scope of this study, but some interesting results emerged from the brief scrutiny of two specific 3-grams in the Organisational and Process categories. At least two features seem to be at play, contributing to the overrepresentation in translated English of (some) 3-grams in these categories:

- Source languages shining through, thus pointing to a strong similarity between German, Norwegian and Swedish in the functional makeup of fiction texts in these categories;
- Translators' very frequent use of a (limited) set of salient expressions that are lexicalised as recurring 3-word sequences within these categories (thus resulting in overrepresentation in terms of overall tokens).

The second point is related to what Baroni and Bernardini (2003: 87) report to have found, namely that there is some "(weak) evidence that there is a systematic difference between translated and original texts in terms of collocational patterns," that is potential universal features. Furthermore, although they admit that the reason for this slight difference could not be fully determined on the basis of their data, they speculate as to "whether such differences are due to a general tendency for translators to use more fixed expressions, or whether there are specific fixed expressions that tend to be favoured by translators (or by original writers)" (Baroni and Bernardini 2003: 87).

Translations from the very closely related languages Norwegian and Swedish did in fact overlap in all the categories in terms of significant vs. non-significant results. The two categories they did not share with translations from German were Comparison and Spatial. Again the source languages' frequent use of such expressions seems to trigger the observed overrepresentation in English translations. This slight difference between translations from Norwegian and Swedish on the one hand and German on the other should be investigated further, possibly with a more complex modelling, and definitely on the basis of more data. For the functional categories that were found to significantly differ between EO and translations from German (Existential and Respect), it was speculated that German generally makes less use of existential constructions than English, thus resulting in a markedly lower use in translations into English, and that the English texts in the corpus may contain more direct speech than the German originals, resulting in an underrepresentation of 3-grams with reporting verbs in translations from German.

The study has some obvious shortcomings that need to be repeated. The selection of source languages is restricted to relatively closely related, and typologically similar, languages. It would therefore be of interest to replicate the study with access to similar material with translations from typologically more different languages. Moreover, the choice of 3-grams as the basis for a functional analysis is far from foolproof and could indeed be expanded in order to be representative of a greater portion of the texts. Nevertheless, this method has proved fruitful in the past and was adopted as a testbed for this expansion of including translations from several languages. Further scrutiny and comparison of more individual 3-gram types would have been interesting to include but has to await further study.

These shortcomings notwithstanding, the study lends some support to the claim that translated language (English in this case) must be considered a "'dialect' of a language unconsciously adopted by translators" (Baroni and Bernardini 2006: 272). A pertinent question arising from this is whether this 'dialect' is universal, in the sense that similar features are favoured in translations from and into several languages. Although the current study has dealt with translated English from several sources it has focused exclusively on translated *English* from a limited set of (Germanic) source languages, thus a general conclusion to this question cannot be drawn. However, I would rather suggest that one of the main results emerging from this study to some extent echoes Teich's findings, suggesting that "[t]he best performing features in our study are those that attest to the 'fingerprints' of the source on the target, what has been called 'source language shining through' (Teich 2003: 113). In Halverson's (2017) terms, it may be argued that the three source languages do in some respects seem to represent a collective gravitational pull in translation into English. However, it should be remembered that, at the level of functionally classified 3-grams, English original and translated texts do behave similarly in the majority of the categories, regardless of source language.

NOTES

[1] R Core Team (2019): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.r-project.org/>.

[2] Following Johansson (2007: 9), comparable corpora are defined here as corpora that "contain original texts in two or more languages matched by criteria such as genre, time of publication, etc." (see also Ebeling and Ebeling 2020).

[3] Johansson, Stig, Ebeling, Jarle and Oksefjell, Signe (1999/2001): The English-Norwegian Parallel Corpus: Manual. *Department of British and American Studies, University of Oslo*. Consulted on 4 April 2022, <http://www.hf.uio.no/ilos/english/services/omc/enpc/ENPCmanual.pdf>.

[4] The English-Norwegian Parallel Corpus is a password protected resource, with restricted access through <https://tekstlab.uio.no/glossa2/omc4> or <http://www.tekstlab.uio.no/cgi-bin/omc/PerlTCE.cgi>.

[5] The English-Swedish Parallel Corpus is a password protected resource, with restricted access through <https://spraakbanken.gu.se/en/resources/espc> or <http://www.tekstlab.uio.no/cgi-bin/omc/PerlTCE.cgi>.

[6] University of Oslo (1999-2008): Oslo Multilingual Corpus - background and use. *University of Oslo.no*. Consulted on 3 April 2022, <https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/index.html>.

[7] The Oslo Multilingual Corpus is a password protected resource, with restricted access through <https://tekstlab.uio.no/glossa2/omc4>.

[8] The word count for English translations from Swedish differs slightly from the number (333,375) reported in Altenberg, Aijmer, *et al.* (2001), and is most likely due to different parameters and programs used for counting.

[9] See Biber, Conrad, *et al.* (2003: 74, 75) for the rationale for this cut-off point in terms of frequency.

[10] Anthony, Laurence (2019): AntConc (version 3.5.8w). Tokyo: Waseda University. Available from <http://www.laurenceanthony.net/>.

[11] For comparison, in the original studies, drawing on English originals and translations from Norwegian only, the combined number of 3-gram types was 1,911.

[12] Cf. Section 4 and Table 6. See also Ebeling and Ebeling (2018) for a detailed outline of how this was done.

[13] The category Respect only contained one 3-gram type (*apart from the*) in the ENPC+ material and was left out of the analysis. This is also the case in the current material and only 14 categories will be part of the analysis.

[14] See Ebeling and Ebeling (2017, 2018) for a more detailed description of how to determine category membership of 3-grams.

[15] The code identifies the corpus from which the example is taken (ENPC+), the author of the text (MoAl = Monica Ali), text number by that author (1) and language (E).

[16] R version 3.2.4.

[17] For the purpose of the current study, effect size will also be added in order to get a better grasp of how strong the relationship between two sets of data is, i.e. the strength of the relationship between the means in each category between English originals and translations from German, Norwegian and Swedish.

[18] Individual 3-gram types within each category will produce more or fewer tokens, relatively speaking, thus contributing more or less to the overall token count within each category. However, as the focus is mainly on functional categories rather than individual items, this is not seen to invalidate the findings.

[19] Apart from five categories in EO (Evaluative, Organisational, Report, Rhematic, Spatial), the data are normally distributed, according to the Shapiro-Whilk normality test. Nonetheless, the *t*-test was chosen due to its general robustness, along with the effect size measure Cohen's *d* (in the "effsize" package in R), but, for the five categories in question, a Wilcoxon rank sum test was also run, with similar results regarding significance and effect size (test statistic *z*). All statistical tests are applied as implemented in R (R version 3.6.2).

[20] All categories with a statistically significant *p*-value (in bold) have a medium or large effect size.

[21] Comparison (ET_NO), Evaluative (both), Organisational (both), Process (ET_NO), Reporting (both), Rhematic (EO), Spatial (both), Thematic Stem (ET_NO).

[22] Similar effect sizes were reported when using Cohen's *d* and standardised test statistic *z* by the square root of the number of pairs (39 in this case) with a medium-large effect size for all the statistically significant results.

23. Evaluative (EO), Organisational (EO), Reporting (both), Rhematic (EO) and Spatial (EO).
24. The texts are aligned at sentence level with the Translation Corpus Aligner (Hofland and Johansson 1998) and are made searchable through the web-based version of the search interface Translation Corpus Explorer (Ebeling 1998).
25. The basic idea of the gravitational pull hypothesis "is that highly salient linguistic items (lexis or grammatical constructions) would be more likely to be chosen by translators and thus be over-represented in translational corpus data" (Halverson 2017: 9).
26. Raw frequencies cannot be compared directly and are only meant to illustrate the ranking of each of the 3-gram types in each of the corpora.
27. The T added to the code, in front of the language (E), identifies the text as a translation, that is Example (3) is from the OMC, and it is an English translation (TE) of text 1 by the author Thomas Bernhard (TBE).
28. The strong relationship between concessive *ändå* and *all the same* is further substantiated by Altenberg (2002: 28), who finds that *all the same* is the fourth most frequent translation of Swedish *ändå* in the English-Swedish Parallel Corpus.
29. GECCo Project (2011-2017): German English Contrasts in Cohesion. *GECCo Project*. Consulted on 3 April 2022, <http://www.gecco.uni-saarland.de/GECCo/index.html>.
30. **Bold**: overlap EO and ET_SW; *Italics*: overlap EO and ET_NO and ET_SW; Light grey shade: overlap EO and ET_GE; Dark grey shade: overlap ET_GE and ET_NO; ***Bold italics***: overlap ET_NO and ET_SW; no highlighting: no overlap with other corpora among the top five.
31. The versatile nature of *way* and sequences including *way* have been thoroughly investigated monolingually by Sinclair (1999) and multilingually, in other words in a translation perspective, by Johansson (2009).

REFERENCES

Altenberg, Bengt (1998): On the phraseology of spoken English: The evidence of recurrent word-combinations. *In*: Anthony Paul Cowie, ed. *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, 101-122.

Altenberg, Bengt (2002): Concessive connectors in English and Swedish. *In*: Hilde Hasselgård, Stig Johansson, Bergljot Behrens and Cathrine Fabricius-Hansen, eds. *Information Structure in a Cross-linguistic Perspective*. Amsterdam: Rodopi, 21-43.

Altenberg, Bengt and Aijmer, Karin (2000): The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies. *In*: Christian Mair and Marianne Hundt, eds. *Corpus Linguistics and Linguistic Theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999.* Amsterdam: Rodopi, 15-33.

Baker, Mona (1993): Corpus linguistics and translation studies: Implications and applications. *In*: Mona Baker, Gill Francis and Elena Tognini-Bonelli, eds. *Text and Technology*. Amsterdam/Philadelphia: Benjamins, 233-250.

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 Bernardini, Silvia (2003): A preliminary analysis of collocational differences in monolingual comparable corpora. *In*: Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery, eds. *UCREL Technical Papers*. (Proceedings of the Corpus Linguistics 2003 conference, Lancaster, 28-31 March 2003). Vol. 16. Lancaster: Lancaster University, 567-586.

Biber, Douglas, Conrad, Susan and Cortes, Viviana (2003): Lexical bundles in speech and writing: An initial taxonomy. *In*: Andrew Wilson, Paul Rayson and Tony McEnery, eds. *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Frankfurt: Peter Lang, 71-105.

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

Culpeper, Jonathan and Kytö, Merja (2002): Lexical bundles in Early Modern English dialogues: A window into the speech-related language of the past. *In*: Teresa Fanego, Belén Méndez-Naya, and Elena Seoane, eds. *Selected Papers from 11 ICEHL*. (Sounds, Words, Texts and Change, Santiago de Compostela, 7-11 September 2000). Vol. 2. Amsterdam: John Benjamins, 45-63.

De Baets, Pauline, Vandevoorde, Lore and De Sutter, Gert (2020): On the usefulness of comparable and parallel corpora for contrastive linguistics. Testing the semantic stability hypothesis. In: Renata Enghels, Bart Defrancq and Marlies Jansegers, eds. *New Approaches to Contrastive Linguistics. Empirical and Methodological Challenges*. Berlin/Boston: De Gruyter, 85-126.

De Sutter, Gert, Goethals, Patrick, Leuschner, Torsten and Vandepitte, Sonia (2012): Towards methodologically more rigorous corpus-based translation studies. *Across Languages and Cultures*. 13(2):137-143.

Ebeling, Jarle (1998): The Translation Corpus Explorer: A browser for parallel texts. *In*: Stig Johansson and Signe Oksefjell, eds. *Corpora and Cross-linguistic Research*. Amsterdam: Rodopi, 101-112.

Ebeling, Jarle and Ebeling, Signe Oksefjell (2013): *Patterns in Contrast*. Amsterdam: Benjamins.

Ebeling, Jarle and Ebeling, Signe Oksefjell (2018): Comparing n-gram-based functional categories in original versus translated texts. *Corpora*. 13(3):347-370.

Ebeling, Signe Oksefjell and Ebeling, Jarle (2017): A functional comparison of recurrent word-combinations in English original vs. translated texts. *ICAME Journal*. 41:31-52.

Ebeling, Signe Oksefjell and Ebeling, Jarle (2020): Contrastive analysis, *tertium comparationis* and corpora. *Nordic Journal of English Studies*. 19(1):97-117.

Fabricius-Hansen, Cathrine (2005): Elusive connectives. A case study on the explicitness dimension of discourse coherence. *Linguistics*. 43(1):17-48.

Frawley, William (1984): Prolegomenon to a theory of translation. *In*: William Frawley, ed. *Translation: Literary, Linguistic and Philosophical Perspectives*. Newark: University of Delaware Press, 159-175.

Granger, Sylviane (1996): From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. *In*: Karin Aijmer, Bengt Altenberg and Mats Johansson, eds. *Papers from a Symposium on Text-based Cross-linguistic Studies*. (Languages in contrast, Lund, 4-5 March 1994). Vol. 88. Lund: Lund University Press, 37-51.

Granger, Sylviane (2018): Tracking the third code. A cross-linguistic corpus-driven approach to metadiscursive markers. *In*: Anna Čermáková and Michaela Mahlberg, eds. *The Corpus Linguistics Discourse. In Honour of Wolfgang Teubert*. Amsterdam: John Benjamins, 185-204.

Halverson, Sandra (2017): Gravitational pull in translation. Testing a revised model: New methodological and theoretical traditions. *In*: Gert de Sutter, Marie-Aude Lefer and Isabelle Delaere, eds. *Empirical Translation Studies. New Methodological and Theoretical Traditions*. Berlin: De Gruyter Mouton, 9-45.

Hofland, Knut and Johansson, Stig (1998): The Translation Corpus Aligner: A program for automatic alignment of parallel texts. *In*: Stig Johansson and Signe Oksefjell, eds. *Corpora and Cross-linguistic Research*. Amsterdam: Rodopi, 87-100.

Johansson, Stig (2002): Towards a multilingual corpus for contrastive analysis and translation studies. *In*: Lars Borin, ed. *Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University*. (Parallel corpora, parallel worlds, Uppsala, 22-23 April 1999). Vol. 43. Uppsala: Brill, 45-59.

Johansson, Stig (2007): *Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam: John Benjamins.

Johansson, Stig (2009): Which way? On English *way* and its translations. *International Journal of Translation*. 21(1-2):15-40.

Johansson, Stig and Hofland, Knut (1994): Towards an English-Norwegian Parallel Corpus. *In*: Udo Fries, Gunnel Tottie and Peter Schneider, eds. *Papers from the Fourteenth*

International Conference on English Language Research on Computerized Corpora. (Creating and Using English Language Corpora, Zurich, 1993). Amsterdam: Rodopi, 25-37.

Kjellmer, Göran (1991): A mint of phrases. *In*: Karin Aijmer and Bengt Altenberg, eds. *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman, 111-127.

Kunz, Kerstin, Degaetano-Ortlieb, Stefania, Lapshinova-Koltunski, Ekaterina, Menzel, Katrin and Steiner Erich (2017): English-German contrast in cohesion and implications for translation. *In*: Gert de Sutter, Marie-Aude Lefer and Isabelle Delaere, eds. *Empirical Translation Studies. New Methodological and Theoretical Traditions*. Berlin: De Gruyter Mouton, 265-312.

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

Lefer, Marie-Aude (2012): Word-formation in translated language: The impact of language-pair specific features and genre variation. *Across Languages and Cultures*. 13(2):145-172.

Mauranen, Anna (1998): Form and sense relations as seen through parallel corpora. *In*: Wolfgang Teubert, Elena Tognini-Bonelli and Norbert Volz, eds. *Proceedings of the Third European Seminar "Translation Equivalence."* (Translation Equivalence, Montecatini Terme, 16-18 October 1997). Mannheim: TELRI, 159-173.

Mauranen, Anna (2000): Strange strings in translated language: A study on corpora. *In*: Maeve Olohan, ed. *Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects*. Manchester: St Jerome, 119-141.

Mauranen, Anna (2007): Universal tendencies in translation. *In*: Gunilla Anderman and Margaret Rogers, eds. *Incorporating Corpora. The Linguist and the Translator*. Clevedon: Multilingual Matters, 32-48.

Moon, Rosamund (1998): *Fixed Expressions and Idioms in English. A Corpus-based Approach*. Oxford: Clarendon Press.

Sinclair, John (1999): A way with common words. *In*: Hilde Hasselgård and Signe Oksefjell, eds. *Out of Corpora. Studies in Honour of Stig Johansson*. Amsterdam: Rodopi, 157-179.

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