Accepted Manuscript version – Edinburgh University Press

Published in *Corpora* 13(3), pp. 347–370

https://www.euppublishing.com/doi/abs/10.3366/cor.2018.0153

Comparing n-gram-based functional categories in original vs. translated texts

Jarle Ebeling and Signe O. Ebeling
University of Oslo

Abstract

The study outlines and tests a method for comparing the use of functional categories consisting of high-frequency 3-grams in original and translated texts. The 3-grams are extracted from a corpus of contemporary English fiction texts (EO) and a comparable corpus of fiction texts translated into English from Norwegian (ET). The two varieties contain the same number of texts, thirty-nine, and approx. the same number of words, 1.3-1.4 million. Several different baselines against which to normalize the 3-gram frequencies are tested and a way of evening out the initial differences between the token counts of EO and ET is proposed. These last two points have an impact on the extent to which some of the categories differ statistically. On the basis of the comparison of the token counts of the 3-grams extracted for the study, it seems as if most differences are a matter of degree, rather than being systemic at the level of the functions investigated.

1 Introduction

Discovering differences and similarities between original and translated texts in the same language is of interest to the general linguist, to the translation scholar and the contrastivist alike, and has several practical applications in applied linguistics. Moreover, translations are great vehicles of cultural dispersion, as pointed out by Halliday: '... translation is a very important process in human life, in part because it is typically expanding the meaning potential of the target language, by bringing in things from outside' (Halliday in Martin (ed.), 2013: 241). However, Halliday adds that 'linguistic evaluation of translations is a very high level of demand to make' (ibid.). In the study reported here, we set out to explore a method for comparing functional categories of 3-grams extracted from texts originally written in English (EO) with texts translated into English from Norwegian (ET). Both sets of texts are made up of contemporary fiction only. The fact that the translated texts are originally written in Norwegian inevitably has an impact on the results emerging from applying the method,[1] but not on the method itself, which is the focus of the study. Our main interest in exploring such a method is to find out whether it can be used to confirm or refute our claim that translations are a good *tertium comparationis* (common ground) in Contrastive Analysis (cf. Ebeling & Ebeling 2013a, ch. 2).

Previous studies of recurrent word combinations (lexical bundles, n-grams) have, among other things, focused on similarities and differences between registers or text types (e.g. Biber *et al*. 1999; 2004, Stubbs & Barth 2003, Simpson-Vlach & Ellis 2010), disciplines (e.g. Cortes 2004, Hyland 2008), learner vs. native-speaker language production (e.g. De Cock 2004, Ädel & Erman 2012, Paquot 2013), languages (e.g. Cortes 2008, Granger 2014, NN in press b) and original vs. translated text in the same language or across languages (e.g. Baker 2004, Wang & Qin 2008, Xiao 2011, Gries & Wulff 2012, Ebeling & Ebeling 2013a, Lee

[1] This is indeed what some translations scholars have found, e.g. Wang and Qin (2008) on original Chinese vs. Chinese translated from English.

2013). The method explored here has taken inspiration from several of the above-mentioned studies, but differs from them in that only 3-grams are included in the study, not e.g. 4-grams (see Biber *et al*. 2004), and without giving any special attention to a particular type of 3-gram from the outset (e.g. Baker 2004, Xiao 2010, 2011).

Another point that sets the current study apart from many of the aforementioned studies, is its focus on an overall classification of all extracted 3-grams into 15 functional categories. This enables us to detect if the differences and similarities we uncover are tied to particular functions as opposed to (only) making claims about differences and similarities in general, or at the level of individual items, between the two "varieties", English original texts (EO) and English translated texts (ET).

The study is structured as follows. In Section 2, we present the functional classification scheme, before we, in Section 3, go on to describe the corpus data (3.1), how the material for the study was extracted and delimited (3.2) and how we normalize the frequencies according to different baselines (3.3). Section 4 deals with the statistical approach to the comparison of the functionally classified 3-grams, while Section 5, briefly, discusses a few points in relation to the outcomes of the statistical analysis. Section 6 offers a conclusion and some thoughts on further research.

Before we delve into the study proper, one could ask whether our EO texts can be considered representative of modern English fiction and thus be used as a good baseline or gold standard with which the translated texts can be compared. This is, of course, not easy to answer, because it begs the question what constitutes a balanced and representative corpus. To address this issue without making it a major point of the study, we compared our EO texts with the fiction part of BNC Baby along the same principles as when we compare EO and ET in our own material, and found few significant differences. Appendix 1 contains the results of the comparison between EO and BNC Baby.

2 Functional classification of 3-grams

It will take us too far afield, and space does not allow us, to survey the many studies that in some way or other are relevant to the current investigation with regard to the way we classify 3-grams into the several functional categories. We refer the interested reader to NN (in press a and b), where some of these studies are described and discussed in more detail. However, the taxonomy used for the classification of 3-grams needs some comments to make the current study comprehensible.

The functional classification draws heavily on Altenberg (1998), Moon (1998) and Biber *et al*. (2004). But, since their classification schemes were primarily developed with other types of texts and sequences (e.g. spoken language, academic genres, fixed expressions and idioms, 4-grams) in mind, some adjustments and additions had to be made. Moreover, some categories emerged as a result of the actual 3-grams encountered in the material. It is beyond the scope of this paper to go into details about our choice of (a mixed) taxonomy, but see NN (in press b) for a more elaborate account of each of them, as well as the rationale for not choosing just one of them.

We operate with fifteen categories in all, twelve of which are informational in nature, following Moon (1998). These are unmarked in the overview given in Table 1, while the additional three categories are marked in bold. One of the categories, Respect, only has one member, *apart from the,* and will not be discussed further. In Table 1 we give a brief definition of each of the functional categories accompanied by 2-3 examples of 3-grams included in the respective categories. The number coming after the example 3-grams shows

the approx. number of 3-gram types in each category. For EO, Modalizing is the largest category with 341 entries, while Spatial is the largest one for ET with 342 entries.[2]

[TABLE 1 GOES HERE]
Table 1. Functional categories of 3-grams in the material

A full discussion of the reasoning behind the choices made when performing this kind of functional classification really warrants a research article, or a research programme, of its own. Having said that, we should note the following, which impinges on the way individual 3-grams were classified. First of all, we did not allow dual membership. This means that some categories had to be given priority when a 3-gram could potentially have belonged to more than one category. For example, the decision was taken that Modalizing trumps all other categories, if there is a modal(izing) item, typically a verb, present.[3] The only exception is Contingency 3-grams introduced by *if*. Similarly, Spatial and Temporal trump other categories, if there is a spatial or temporal element present. Moreover, Contingency 3-grams, many with *if*, also trump e.g. Existential. Fortunately, few conflicts between the categories were encountered and most could be resolved by looking up the potentially ambiguous 3-grams in their context in the corpus. In true cases of ambiguity, the most frequent function of a 3-gram in context was decisive for category membership. A case in point is *back to the*, where it in example (1) could arguably be said to belong to the Temporal category, and in example (2) to the Spatial category. For the purpose of this study, *back to the* is classified as Spatial because this is its most frequent function by far in our material (171 occurrences out of 173).

(1) ... Jay cupped his hand and moved him back to the beginning. [MoAl1E]

(2) Crutch in hand, she walks back to the bed and sits down beside me. [PaAu1E]

3 The corpus and the material extracted for the study

3.1 The corpus

This study draws on material from the English-Norwegian Parallel Corpus+ (ENPC+), a balanced, bidirectional translation corpus containing thirty-nine original fictional texts in English (EO) and thirty-nine in Norwegian (NO) and their respective translations (NT, ET). The original ENPC contains extracts of books of between 10,000 and 17,000 words (Johansson *et al*., 1999/2001), while the ENPC+ contains an additional nine English and nine Norwegian original texts, which are whole books of between 52,000-211,000 words. Each subcorpus amounts to around 1.3 million (EO, NO, NT) and 1.4 million (ET) words (see Ebeling & Ebeling 2013a: 84*ff* for more detailed information about the ENPC+).

As the investigation focuses on intra-linguistic features of English, data will only be culled from the English originals and English translations subcorpora. The translational and non-translational components are comparable in the sense that both pools of texts can be broadly defined as contemporary fiction, covering the same time period from the 1980s to 2012, as well as containing a similar amount of text (in number of texts as well as running words).

3.2 Data delimitation and extraction

[2] When reporting type frequencies, it should be kept in mind that they are sensitive to text size, although more so for words than for high-frequency 3-grams. See Biber (2006: 253) for a discussion of this point.
[3] In addition to modal auxiliaries, verbs expressing attitude, possibility/probability or certainty towards a proposition, e.g. *be going to*, *know*, *think*, *want to*, *perhaps* and *seem*, are also included in this category.

Most previous studies dealing with functional classes of lexical bundles and n-grams have focused on 4-word sequences. 4-grams seem to fall more clearly into functional groupings than 3-grams. Indeed, Hyland (2008: 8) 'decided to focus on 4-word bundles because they are far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles'. As our corpus is relatively small, we have chosen to focus on 3-grams to get a larger 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 be part of the same 4-gram. Moreover, as pointed out by Culpeper and Kytö, 3-word combinations ' offer a good compromise between the great number of different two-word combinations and the small number of different four-word combinations' (2002: 45).

To counter the potential effect of topic-specific items and/or idiosyncratic uses by the individual authors or translators we require the 3-grams to occur in at least twenty-five per cent of the texts, i.e. in ten of the thirty-nine texts. Another reason for this relatively strict requirement of distribution across texts has to do with the fact that nine of the texts in each subcorpus are substantially longer than the rest; thus, this will ensure that the 3-grams also have to occur in at least one of the shorter texts.[4] We also introduced an additional, and quite conservative, threshold requiring each 3-gram to occur with a frequency of at least twenty pmw, i.e. twenty-six and twenty-eight times respectively in each of the subcorpora (EO, ET). 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). We refer to these two conditions, dispersion and recurrence, collectively as **the threshold**. It should be mentioned here 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 subcorpora, this does not mean that it was not attested at all in that subcorpus. In fact all the 3-grams that reached the threshold in EO are attested in ET and vice versa. This is something we shall explore when we perform the statistical analyses.

The 3-gram types were extracted using AntConc (Anthony, 2014), where, in addition to setting the above-mentioned threshold, we also made some changes to the default settings to ensure that: (1) tags/mark-up is not part of the 3-grams; (2) apostrophe and hyphen are not treated as word-delimiters, e.g. *n't* is counted as one word and not two (this will not apply to contracted PRON+VERB forms, such as *I'm*, since these have already been split (*I 'm*) in the corpus files.); (3) 3-grams do not cross s-unit (sentence) boundaries.[5] By default, AntConc allows n-grams to run across commas, colons and semi-colons (i.e. the program ignores them when creating n-gram sequences); we chose to do the same. As it turns out, this has relatively little bearing on the number of 3-grams extracted for our study, as the high-frequency 3-grams tend not to contain punctuation marks anyway.

The method of extraction gave us two comparable lists of 3-gram types, one for the EO texts and one for the ET texts, amounting to 1,408 and 1,468 3-gram types respectively. These were then classified into the 15 functional categories that form the basis for the comparison between the EO and the ET subcorpora. The actual numbers used when applying the statistical measures are the token counts for each category, i.e. the number of times we encountered a 3-gram type belonging to that category. We counted the tokens by having a Perl script read each text and compare every 3-gram in that text with the 3-grams in the type lists. This gave us two matrices, one for EO and one for ET, with thirty-nine rows (one for

[4] Admittedly, this is a rather basic way of dealing with the complex question of dispersion (cf. Gries, 2008 and Lijffijt & Gries, 2012), but at least we ensure that a handful of (the longer) texts in our material does not make up the total pool of recurrent 3-grams.

[5] This meant that the following changes were made to the Global Settings in AntConc: Tags: hide tags and Token definition: User-defined token class; Append following definition: '- for the first two restrictions, while for the third, the following change was made to the Tool Preferences: Untick the Replace linebreaks box.

each text) and fifteen columns (one for each functional category). The Perl program also counted the total number of words in each text, the total number of 3-grams, the number of s-units (orthographic sentences) and the total number of threshold tokens, i.e. the total number of tokens for all fifteen categories for that text that meet the threshold. We will use these overall frequencies of words, 3-grams, s-units and total threshold tokens as baselines to create different normalized frequencies. These will in turn be used as input to the statistical tests.

At this point, and as a backdrop to the more fine-grained statistical analysis, it can be observed that there is a marked difference between the subcorpora when we compare the token counts, EO: 83,827 vs. ET: 87,878, but not when we compare the type counts, EO: 1,408 vs. ET: 1,468.[6]

Figure 1 shows the proportional distribution (in per cent) of 3-gram tokens in EO and ET for the most frequent categories.

[FIGURE 1 GOES HERE]

Figure 1. Proportional distribution of 3-gram tokens in EO and ET (in per cent)

The categories Modalizing and Spatial stand out. There are clearly proportionally more Modalizing tokens in the EO texts compared with the other categories, while there are more Spatial tokens in the ET texts. EO also has, proportionally, more Thematic stems ('thematic' in Figure 1) than ET has, while the opposite is the case for e.g. Fragment.

3.3 Normalization of frequency counts

After having classified all 3-grams above the threshold in the two subcorpora, EO and ET, we counted the number of token occurrences of each functional category for each text. As mentioned, we also counted the number of words, 3-grams, s-units and overall token occurrence above the threshold for each text. By "overall token occurrence above the threshold" is meant all 3-gram tokens regardless of functional category which meet the threshold, i.e. tokens resulting from the 1,408 and 1,468 type lists respectively. Table 2 shows the counts for four of the texts,[7] two from EO and two from ET, for two of the functional categories, viz. Fragment and Modalizing.

Table 2. Token counts and normalized frequencies for four texts

[TABLE 2 GOES HERE]

The important numbers in Table 2 are the normalized frequencies, where the raw frequencies have been divided by number of words, 3-grams, s-units and total number of threshold tokens. To get more human-readable numbers, we multiply words and 3-grams per 1,000 instances and s-units and tokens per 100 instances. These are the numbers we run the statistical tests on. For instance, for the text MoAl1E, the number of Modalizing tokens is 1,058. This number in turn is divided by 77,364 (# of words), 63,068 (# of 3-grams), 7,218 (# of s-units) and 4,731 (# of threshold tokens). Normalized by 1,000 (words, 3-grams) and 100 (s-units, 3-grams threshold tokens), this gives the frequencies 13.68, 16.78, 14.66 and 22.36. Such normalized frequencies are calculated for all the 78 texts.

4. Test statistics

We have run two statistical tests on the material, an independent, two-tailed *t*-test with Welch's correction and the Mann-Whitney U test, also known as the Wilcoxon rank-sum

[6] The Log-likelihood calculator provided by Paul Rayson (http://ucrel.lancs.ac.uk/llwizard.html) shows a statistically significant difference between the token counts in the two subcorpora, but not between the type counts.

[7] See Ebeling & Ebeling (2013a) for an overview of the texts and text codes in the corpora.

test.[8] In both cases we apply the tests as they are implemented in R.[9] In the following, we will only report the *p*-values from the *t*-test, since both tests showed the same tendencies, although the *p*-values output from the two different tests were not, of course, identical. The reason for applying both tests is that not all the categories are normally distributed, which is an underlying assumption of the *t*-test. On the other hand, the *t*-test is considered more robust than the Mann-Whitney U test, hence we chose to use this as the basic test statistics.

If the test shows a significant effect, the *p*-value will be <= 0.05, else > 0.05 at the 95% confidence level with DF = seventy-six. Note that the test does not say anything about whether a specific functional category is more or less frequent in EO or ET.

Before we get too caught up in the results of the statistical measures, it must be stressed that the *p*-values are, for us, primarily seen as a good way into the more interesting qualitative investigation of the data; why, for instance, is such and such a category more/less frequent in EO than in ET? Is it source language shining through, target language normalization, or something entirely different, which would somehow invalidate our claim that translations are a good *tertium comparationis* when doing Contrastive Analysis?

Table 3 shows the *p*-values resulting from the *t*-test when applied to the normalized 3-gram token counts of the 14 categories we have counted (excluding Respect).

Table 3. *p*-values calculated against four different baselines: words, 3-grams, s-units and threshold tokens

[TABLE 3 GOES HERE]

What we can note right away is that for Existential, Quantifying/Intensifying and Thematic stem the variety does not have a significant impact, no matter which baseline we apply. For the categories Contingency and Rhematic, there are interesting differences in that variety has an impact when we use threshold tokens as baseline, but not when we use words, 3-grams or s-units.

With regard to the baselines chosen, words, all 3-grams, s-units or all tokens above the threshold regardless of category, we find that words is perhaps the least appropriate, since a 3-gram spans more than one word. We have earlier (Ebeling & Ebeling, 2013a) argued for the use of s-units as a good baseline, but if we normalize by s-unit, we throw ourselves at the mercy of the style of the author in terms of length of s-units.[10]

3-gram tokens seem to be a reasonable baseline, if not ideal, since we have texts of very different lengths and need to normalize the number of tokens by some measure that takes text length into account. Note incidentally that all s-units with less than 3 words are skipped when we count 3-gram tokens.

As regards the use of all tokens above the threshold regardless of functional category as baseline, the reasoning is that since we have already reduced the potential of the types of 3-grams that will be counted, i.e. those occurring above the threshold (twenty pmw and in twenty-five per cent of the texts), we should divide the actual number of tokens found for a category in a particular text against that potential, which is the number of tokens of all the types above the threshold (in that text). On the other hand, the overall frequency of threshold tokens in a text varies quite a lot and is not as predictable across texts as is the total number of 3-grams. In one text, for instance, the total number of threshold tokens accounts for eight per

[8] See Lijffit *et al*. (2014) for a discussion of the use of these and other tests when comparing word frequencies in corpora.

[9] R version 3.2.3; RStudio version 0.99.491.

[10] In theory we are also at the mercy of the style of the translator, but earlier research (Ebeling & Ebeling 2013b) has shown that translators mainly translate sentence by sentence when translating between English and Norwegian, i.e. there is mostly a one-to-one correspondence between source and target at the sentence level.

cent of all 3-grams, while for another, the number is down to 5.9%.[11] This means that if one divides the number of threshold tokens for a particular category against the total number of threshold tokens for that text, the normalized frequencies can be very different. The overall picture seems to be that there are more significant differences between EO and ET when normalized frequencies of this last kind are used than when, e.g., words or all 3-grams are used (compare Tables 2 and 4). In what follows, we will only report *p*-values for tokens divided by all 3-grams and all tokens above the threshold (3rd and 5th column in Table 3).[12]

Before interpreting the results of the above tests and investigating the data qualitatively, we wanted to even out the frequency counts by including the token counts for the 3-gram types that did not initially reach the threshold in either EO or ET. As mentioned above, we noticed during the initial extraction of the 3-gram types that all the types attested in the EO texts that reached our threshold were indeed attested in the ET texts and vice versa, even though they did not reach the threshold in the respective subcorpora, either by not occurring at least twenty times per million words or not being attested in at least twenty-five per cent of the texts. Table 4 shows six 3-gram types of the Modalizing category. Two of the 3-grams reach the threshold in both corpora, viz. *are you going* and *be able to*, while the remaining only reach the threshold in either EO or ET. Remember that a 3-gram had to occur at least 26 times in EO and 28 in ET to reach a frequency of 20 pmw.

Table 4. Adding token counts for 3-gram types that initially did not reach the threshold
[TABLE 4 GOES HERE]
The shaded cells in Table 4 show how we included token counts for 3-gram types that did not initially meet the thresholds for either EO or ET. For, e.g., *as i could*, only one more attested occurrence would have meant that the Modalizing category would have included twenty-eight more instances in the ET corpus distributed over the thirty-nine texts.

Figure 2 shows the adjusted proportional distribution of categories, and clearly shows how the differences between EO and ET have been evened out when compared to Figure 1.
[FIGURE 2 GOES HERE]
Figure 2. Adjusted proportional distribution of 3-gram tokens in EO and ET

Next, we shall see how this way of topping-up the token counts affects the *p*-values reported. It can be argued that this way of evening out the initial extraction of 3-gram types obscures interesting differences in the use of individual 3-grams. However, since our main concern here is differences and similarities at the level of the functional categories, we think it is a defensible approach. Moreover, this is an exploratory study testing several ways of comparing original and translated English.

By adding token counts for 3-gram types that initially did not reach the threshold in one of the subcorpora, as we have done in Table 5, we have evened out the initial difference between the EO and the ET texts. This results in non-significant *p*-values for the categories Evaluative and Modalizing, when we normalize by 3-grams (compare Tables 3 and 5).

Table 5. *p*-values calculated against two different baselines based on topped-up token counts: 3-grams and threshold tokens
[TABLE 5 GOES HERE]
Similarly, it has led to the result that Contingency, Fragment and Process do not show significant values when we normalize by threshold tokens. More surprising, perhaps, is the

[11] The number of 3-gram tokens above the threshold accounts for 7.5% of all 3-grams in EO and 7.3% of all 3-grams in ET.
[12] Ideally, one would have wanted a parsed corpus and divided the number of 3-grams by the number of (all kinds of) clauses, as this would have taken into account difference in sentence length.

fact that Thematic stem goes from being non-significant to being significant, when we use threshold tokens as baseline.

Before we go on to investigate what lies behind some of these changes, we should note that the minor categories Organizational, Process and Report have very few 3-gram type entries in EO/ET overall: Organizational has four / eight respectively, Process has ten / seventeen and Report has eighteen / ten. This means that the *p*-values reported for these categories rely on fewer observations in the form of the number of tokens counted for each category in each text. It follows from this that the outcome of the statistical tests are less reliable.

Since the focus of this study is on method and not on the linguistic causes why a particular category is more or less used in one of the varieties, we will restrict ourselves to a few comments regarding the significant *p*-values reported for Comparison, Fragment, Spatial and Temporal.[13] We shall also take a closer look at the Thematic stem category, which shows quite an intriguing difference in *p*-values, 0.9913 vs. 0.0291, depending on the baseline chosen, as shown in Table 5.

5. Findings and discussion

In this section we take a peek behind the scenes, so to speak, and pinpoint the actual 3-grams and frequencies that result in the significant *p*-values reported in Table 5. We will do this by sorting the 3-gram token counts by the difference in the number of tokens between EO and ET. This gives us a good indication of exactly what it is that causes the differences that give rise to the significant *p*-values. We shall illustrate this procedure with the category Temporal (5.1), since similar underlying tendencies seem to be at play in all the four above-mentioned categories (Comparison, Fragment, Spatial and Temporal).

Note that the token counts used here are the ones where we have topped-up the token counts for EO or ET, so that there are equally many 3-gram types for the two varieties. More importantly, the numbers listed in Table 6 are the raw token frequencies for Temporal across all texts, which means that since the EO subcorpus is smaller than the ET subcorpus by 100,000 words, the differences are inflated by the relative difference in size between the two subcorpora, i.e. approx. 7.6%.[14]

In 5.2, we look closer at Thematic stem and why this category seems to vacillate between being significant and non-significant.

5.1 Temporal

The adjusted proportional distribution for the Temporal category in Figure 2 shows that it accounts for roughly the same proportion of all the 3-grams in EO and ET. Still, the *p*-value has consistently pointed to a significant impact of the variety on the dependent variable, i.e. the token counts for each text. The reasons for this can be gleaned from Table 6, where we show the top fifteen 3-grams sorted by their difference in frequency.

Table 6. Temporal 3-grams sorted by difference in number of tokens (raw frequencies)
[TABLE 6 GOES HERE]

[13] Fragment does not show a significate *p*-value, 0.1078, in Table 4, when all threshold tokens are used as baseline, and the combined 3-gram type list is used as the basis for the token counts. However, when the two separate 3-gram type lists were used, it does show a significant *p*-value of 0.0033 (Table 2). For Thematic stem the opposite is the case when all threshold tokens are used as baseline with the combined 3-gram type list. That category was not significant in Table 2 (*p*-value = 0.682), but is significant in Table 4 (*p*-value = 0.0291).

[14] These raw frequencies should not be confused with the normalized frequencies used in the statistical model to calculate the *p*-values.

The difference (Diff.) between EO and ET for, e.g. the 3-gram *a long time* is 167 (295 - 128), for *at the same* 132 and so on. The actual frequencies giving rise to the difference are listed in the Freq. columns. The consistently higher number of tokens in ET compared to EO gives rise to the significant *p*-value for Temporal. Similar differences in token counts are noted for Comparison, Fragment and Spatial.

It would be interesting to go into more detail and investigate the underlying linguistic and/or cultural reasons for these differences by performing a Contrastive Analysis of the translations and the sources, but this lies outside the scope of this paper (see NN in press a and b for a discussion of such issues).

5.2 Thematic stem

Table 3 shows that the choice of baseline has an effect in the sense that different *p*-values are reported for the various categories. However, using words, 3-grams or s-units does not alter the significance level of the *p*-value for any of the categories, but using threshold tokens does, as can be seen in the case of Contingency and Rhematic. In Table 5, Thematic stem also shows a significant difference when threshold tokens are used as baseline, even if this was not the case in Table 3. Thus, the use of the total pool of tokens above the threshold for each text as baseline results in several more significant *p*-values. This can be illustrated if we plot the normalized frequencies for 3-grams and threshold tokens and place the plots the side by side, as in Figure 3.

[FIGURE 3 GOES HERE]

Figure 3. Boxplots of Thematic stem with two different baselines

Note that the Y axes are not directly comparable; nevertheless we see marked differences in the size of the boxes and the overlap of the notches, and we get an additional outlier among the ET texts when threshold tokens are used as baseline. The underlying, text-internal causes for the differences escape us at present, but the reason why the two different baselines produce different results seems to be directly linked to the fact that the total pool of threshold tokens counted for each text, when used to produce a normalized frequency, shows more differences between the texts within each variety, ultimately resulting in a significant difference between EO and ET. Another way of visualizing this is by comparing the ranking of the normalized frequencies for 3-gram and threshold-token baselines respectively. Table 7 shows a snapshot of the ranking of the normalized frequencies, and the sum of the ranking.

Table 7. Ranking of normalized frequencies for Thematic stem based on two different baselines

[TABLE 7 GOES HERE]

The sum of the ranks, 1,590 vs 1,491 for 3-grams and 1,796 vs. 1,285 for threshold tokens, strongly indicates the difference between the varieties for thresholds tokens. This is also noticeable when one compares the ranking of the bottom 10 ranks for this normalized frequency, as most of them belong to EO. When 3-grams are used as baseline, the ranking is much more evenly distributed between EO and ET. The reason for this difference in ranking seems to be that the translations contain more of the threshold tokens overall. This suggests that 'the general rule that frequent items occur even more frequently in translation' (Mauranen 2000: 10) also applies to our Thematic stem category, despite the very conservative threshold we set. Thus, when the number of Thematic stem tokens per text is divided by the total number of threshold tokens, this yields a lower number compared to when Thematic stem tokens are divided by the number of 3-grams per text. This then, seems to hold for enough ET texts to produce a significant *p*-value when the two subcorpora are compared. Why this is the case is both interesting and intriguing, but investigating the causes does, unfortunately, lie

outside the scope of the current study. The differences reported here underline, again, the importance of choosing the most appropriate baseline.

5.3 Summary of findings

The discussion of the frequencies underlying the significant *p*-value for the category Temporal, and by association the categories Comparison, Fragment and Spatial, has revealed that the translation subcorpus, ET, contains more token occurrences of the 3-gram types than EO, also when we take into account that ET contains approx. 100,000 more words than EO.

The study has shown, however, that such token differences do not affect all functional categories, at least not for English texts translated from Norwegian. It follows from this that a careful classification of n-grams is useful, if not a pre-requisite, if the purpose of the study is to make claims about similarities and differences between original and translated texts.

With regard to the category Thematic stem, we investigated why this category in particular was more sensitive to the choice of baseline than the other categories. The underlying reason for this seems to be that a number of the translations, for this category, contain more of the threshold tokens overall resulting in enough lower normalized frequencies to create a significant difference in *p*-value, as shown by comparing the ranks of the two rightmost columns of Table 7.

6 Conclusion

The method developed and tested in this paper addresses the question 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 good knowledge of the second. To pave the way for the quantitative part of the method, all frequent 3-grams in the two subcorpora were classified into 15 functional categories. This gave us the possibility of distinguishing between categories that do and categories that do not show significant differences between the EO and the ET texts.

As part of the methodological approach we employed the *t*-test (and the Mann-Whitney U test) as a kind of litmus test in order to decide (quantitatively) which functional categories that are interesting and meaningful to investigate qualitatively. An important part of the quantitative analysis was the selection of an appropriate baseline and how to adjust for 3-grams that did not reach the initial threshold set for inclusion in the material. Moreover, dispersion and recurrence were taken into account to make provisions for idiosyncrasies that arise when texts of different sizes and compositions are used as part of the comparison.

Our belief that translations are a good *tertium comparationis* is not overturned as a result of this study, since non-significant differences between EO and ET are found in roughly half of the 14 functional categories, depending on the baseline chosen (see Table 5). However, the fact that the other half shows significant results paints a rather complex picture, calling for caution when using translations as the basis for contrastive analysis. Thus, more than anything, we need to take care when using translations, since in some cases we get a skewing effect when source language shines through and when the translators select (unconsciously) a default target language rendering. This skewing effect was quite noticeable for several of the categories, which showed a consistently significant difference between EO and ET. The fact remains, however, that all 3-gram types meeting the threshold in EO were attested in ET and vice versa.

Appendix 1: BNC Baby – fiction vs. EO[15]

[15] http://www.natcorp.ox.ac.uk/corpus/baby/manual.pdf

BNC Baby, which is a four million word sampling of the 100 million-word British National Corpus, most likely needs no detailed introduction. This is what Burnard (ed., 2008) writes about the 'fiction' part of the BNC Baby corpus:

> Texts for the fiction component were selected from texts classified as 'written imaginative', published as books between 1985-1994, as having been produced for an adult audience, and having the genre label W fict prose. From this set of 356 texts, a random sample of about one million words (25 texts) was drawn. The sample was checked to ensure no more than one title by any particular author was selected.

We performed the same exercise on the 25 BNC Baby texts as we did with EO (and ET), that is, we extracted recurrent 3-grams from ten per cent of the texts with the condition that the 3-gram had to occur 20 times per million words. We opted for ten per cent of the texts rather than twenty-five per cent since we had fewer texts overall.

Next we classified the 3-grams according to the same criteria and into the same fifteen functional categories as for EO (and ET). Since the fiction part of BNC Baby only contains twenty-five texts, we reduced the number of EO texts from thirty-nine to twenty-five, making sure to keep both long and short texts among the twenty-five EO texts. The BNC Baby texts are all between approx. 30,000 and 50,000 words long. Table A1 shows the result of the comparison.

Table A1. Comparing EO and BNC Baby on the basis of *p*-values
TABLE A1 GOES HERE]

There are some differences between EO and BNC Baby that would have been interesting to follow up, e.g. the differences for the category Fragment when using the separate lists of 3-gram types. Fragment is possibly the most informational of the categories and could point to differences in content between the two corpora. EO contains more crime fiction than BNC Baby for instance. Following up on this is, however, not part of this study, and will have to await further study.

The overall picture that emerges from the *p*-values for all the functional categories is that EO and BNC Baby do not seem to differ to the extent that some of the categories for EO and ET do. This gives us confidence that our EO texts can indeed be used as a reference point for the comparison with the ET texts to tell us in what ways English translations from Norwegian are different from, or similar to, comparable texts originally written in English for the fifteen functional categories.

Acknowledgements
We would like to thank two anonymous reviewers for their valuable comments and suggestions on a previous version of this paper.

References

Ädel, A. and B. 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, pp. 81–92.

Altenberg, B. 1998. 'On the phraseology of spoken English: The evidence of recurrent word-combinations' in A.P. Cowie (ed.) *Phraseology: Theory, Analysis and Applications*, pp. 101–122. Oxford: Oxford University Press.

Anthony, L. 2014. AntConc (Version 3.4.4w) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/.

Baker, M. 2004. 'A corpus-based view of similarity and difference in translation', *International Journal of Corpus Linguistics* 9 (2), pp. 167–194.
Biber, D. 2006. *University Language. A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins. DOI: 10.1075/scl.23
Biber, D., S. Conrad and V. 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*, pp. 71–105. Frankfurt: Peter Lang.
Biber, D., S. Conrad and V. Cortes. 2004. '"If you look at…": Lexical bundles in university teaching and textbooks', *Applied Linguistics* 25 (3), pp. 371–405.
Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.
Burnard, L. (ed.). 2008. *Reference Guide to BNC Baby (second edition*). http://www.natcorp.ox.ac.uk/corpus/baby/manual.pdf
Cortes, V. 2004. **'**Lexical bundles in published and student disciplinary writing: Examples from history and biology', *English for Specific Purposes* 23, pp. 397–323.
Cortes, V. 2008. 'A comparative analysis of lexical bundles in academic history writing in English and Spanish', *Corpora* 3 (1), pp. 43–58.
Cortes, V. 2015. 'Situating lexical bundles in the formulaic language spectrum' in V. Cortes and E. Csomay (eds), *Corpus-based research in Applied Linguistics. Studies in Honour of Doug Biber*, pp. 197–216. Amsterdam: John Benjamins.
Culpeper, J. and M. Kytö. 2002. **'**Lexical Bundles in Early Modern English dialogues: A window into the speech-related language of the past'. In T. Fanego, B. Méndez-Naya, and E. Seoane (eds), *Sounds, Words, Texts and Change. Selected Papers from 11 ICEHL, Santiago de Compostela, 7–11 September 2000*, pp. 45–63. Amsterdam: John Benjamins.
De Cock, S. 2004. 'Preferred sequences of words in NS and NNS speech', *Belgian Journal of English Language and Literature (BELL)* New Series 2, pp. 225–246.
Ebeling, J. and S.O. Ebeling. 2013a. *Patterns in Contrast*. Amsterdam: John Benjamins.
Ebeling, S.O. and J. Ebeling. 2013b. 'From Babylon to Bergen: On the usefulness of aligned texts', *BeLLs* 3 (1). DOI: http://dx.doi.org/10.15845/bells.v3i1.359.
Granger, S. 2014. 'A lexical bundle approach to comparing languages: Stems in English and French', *Languages in Contrast*, 14:1, 58–72. DOI:10.1075/lic.14.1.04gra. Available from http://hdl.handle.net/2078.1/145325.
Gries, S.Th. 2008. 'Dispersions and adjusted frequencies in corpora', *International Journal of Corpus Linguistics* 13 (4), pp. 403–437.
Gries, S.Th. and Stefanie Wulff. 2012. 'Regression analysis in translation studies' in M.P. Oakes and J. Meng (eds), *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research*, pp. 35–52. Amsterdam: John Benjamins.
Hyland, K. 2008. '"As can be seen". Lexical bundles and disciplinary variation', *English for Specific Purposes* 27, pp. 4–21.
Johansson, S., J. Ebeling and S. Oksefjell. 1999/2001. *The English-Norwegian Parallel Corpus: Manual*. Institutt for britiske og amerikanske studier, Universitetet i Oslo. Available from http://www.hf.uio.no/ilos/english/services/omc/enpc/ENPCmanual.pdf.
Lee, C. 2013. 'Using lexical bundle analysis as discovery tool for corpus-based translation research', *Perspectives* 21 (3), pp. 378–395.
Levshina, N. 2015. *How to do Linguistics with R*. Amsterdam: John Benjamins.
Lijffijt, J. and S.Th. Gries. 2012. 'Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora", International Journal of Corpus Linguistics', *International Journal of Corpus Linguistics* 17 (1), pp. 147-149.

Lijffijt, J., T. Nevalainen, T. Säily, P. Papapetrou, K. Puolamäki and H. Mannila. 2014. 'Significance testing of word frequencies in corpora', *Digital Scholarship in the Humanities* 28 (1). DOI: http://dx.doi.org/10.1093/llc/fqu064

Martin, J.R. 2013. *Interviews with M.A.K. Halliday. Language Turned Back on Himself.* London: Bloomsbury.

Mauranen, A. 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*, pp. 119–141. Manchester: St Jerome.

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

NN. In press a. 'A functional comparison of recurrent word-combinations in English original vs. translated texts', *ICAME Journal* 41 (2017), 31–52. DOI: 10.1515/icame-2017-0002.

NN. In press b. '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.

Paquot, M. 2013. 'Lexical bundles and L1 transfer effects'. *International Journal of Corpus Linguistics* 18(3), 391–417.

R Core Team 2015. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/.

Simpson-Vlach, R. and N.C. Ellis. 2010. 'An academic formulas list: New methods in phraseology research'. *Applied Linguistics* 31(4), 487-512.

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

Teich, E. 2003. *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton de Gruyter.

Wang, K. and H. Qin. 2008. 'A Parallel Corpus-based Study of Translational Chinese'. http://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Wang_and_Qin.pdf

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

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

Table 1. 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*) <= 25
Contingency	expresses a condition, reason, cause or concession.	(*because it was, if he 'd, why did you*) 50-100
Evaluative	similar to modalizing but typically contains an evaluative**Error! Bookmark not defined.** adjective or adverb instead of a verb.	(*'s a good, i 'm sure, just do n't*) EO: 51, ET: 30
Existential	contains existential *there.*	(*and there 's, there were no*) <= 25
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*) 50-100
Modalizing	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*) > 100
Organizational	contains items that are clearly recognizable as text structuring devices.	*all the same, in any case*) <= 25
Process	is represented by manner and means expressions.	*in a way, the way you*) <= 25
Quantifying / Intensifying[16]	contains quantifying and intensifying expressions.	*a glass of, more or less, lot of time*) 50-100
Reporting	includes a reporting verb.	*he said and, no he said*) <= 25
Respect	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*) 1
Rhematic	typically includes a verb followed by (part of) a noun phrase (i.e. the beginning of an object or complement/predicative).	*'s not a, he told me, to give him*) > 100
Spatial	includes a clear spatial reference.	*across the table, back in the, to be there*) > 100
Temporal	includes a clear temporal reference.	*a few days, at the moment, he 'd never*) 50-100
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*) > 100

[16] Even if intensifiers often border on modalizing we have chosen to follow Altenberg (1998) in operating with one category for Quantifying and Intensifying 3-grams.

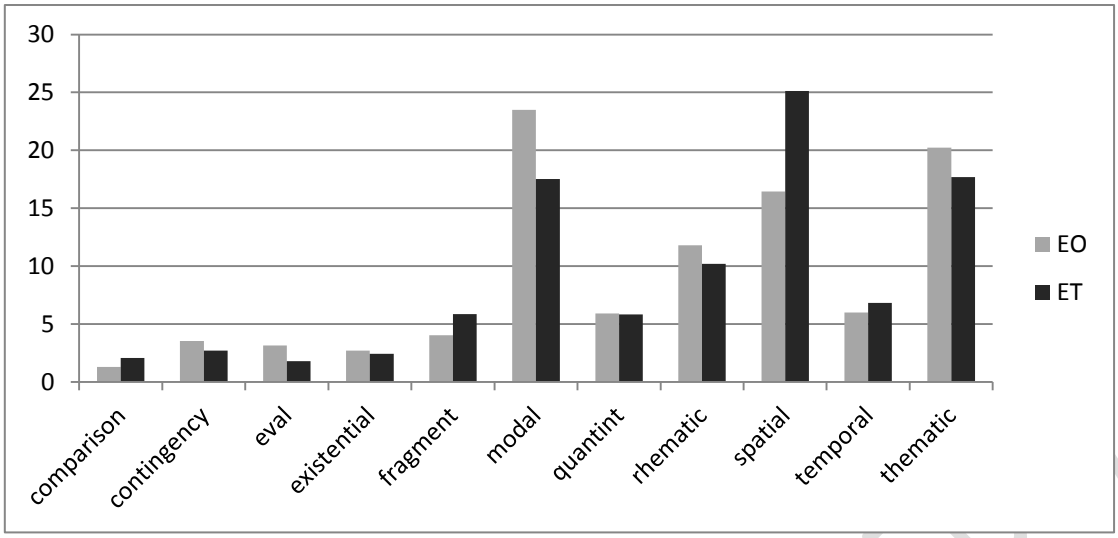

Figure 1. Proportional distribution of 3-gram tokens in EO and ET (in per cent)

Table 2. Token counts and normalized frequencies for four texts

	EO		ET	
	MoAl1E	MW1E	JoNe1TE	JW1TE
# of words	77,364	10,828	138,849	13,042
# of 3-grams	63,068	8,878	113,843	11,326
# of s-units	7,218	982	12,765	858
# of threshold tokens	4,731	636	9,261	700
Fragment tokens	180	15	524	50
Modalizing tokens	1,058	158	1,774	104
Normalized frequencies	Fragment / Modalizing		Fragment / Modalizing	
Tokens / words * 1,000	2.33 / 13.68	1.39 / 14.59	3.77 / 12.78	3.83 / 7.97
Tokens / 3-grams * 1000	2.85 / 16.78	1.69 / 17.8	4.6 / 15.58	4.41 / 9.18
Tokens / s-unit * 100	2.49 / 14.66	1.53 / 16.09	4.1 / 13.9	5.83 / 12.12
Tokens / th. tokens * 100	3.8 / 22.36	2.36 / 24.84	5.66 / 19.16	7.14 / 14.86

Table 3. *p*-values calculated against four different baselines: words, 3-grams, s-units and threshold tokens

Category	tokens / words	tokens / all 3-grams	tokens / s-units	tokens / threshold tokens
Comparison	<0.001	<0.001	0.0035	0.0014
Contingency	0.1426	0.1821	0.1634	<0.001
Evaluative	0.0044	0.0065	0.0026	<0.001
Existential	0.7393	0.7487	0.9445	0.0628
Fragment	<0.001	<0.001	0.0013	0.0033
Modalizing	0.0258	0.0417	0.0145	<0.001
Organizational	<0.001	<0.001	0.0028	<0.001
Process	<0.001	<0.001	<0.001	<0.001
Quantifying/Intensifying	0.4119	0.3737	0.6480	0.6213
Reporting	0.0023	0.0029	0.0059	<0.001
Rhematic	0.3161	0.4063	0.2582	<0.001
Spatial	<0.001	<0.001	<0.001	<0.001
Temporal	<0.001	<0.001	0.0015	<0.001
Thematic stem	0.8522	0.7808	0.9755	0.0682

Table 4. Adding token counts for 3-gram types that initially did not reach the threshold

3-gram	EO freq.	EO dist.	ET freq.	ET dist.
are you going	47	>= 10	48	>= 10
as i can	15	10	28	>= 10
as i could	34	>= 10	27	11
as i know	21	8	30	>= 10
as she could	26	>= 10	18	9
be able to	160	>= 10	189	>= 10

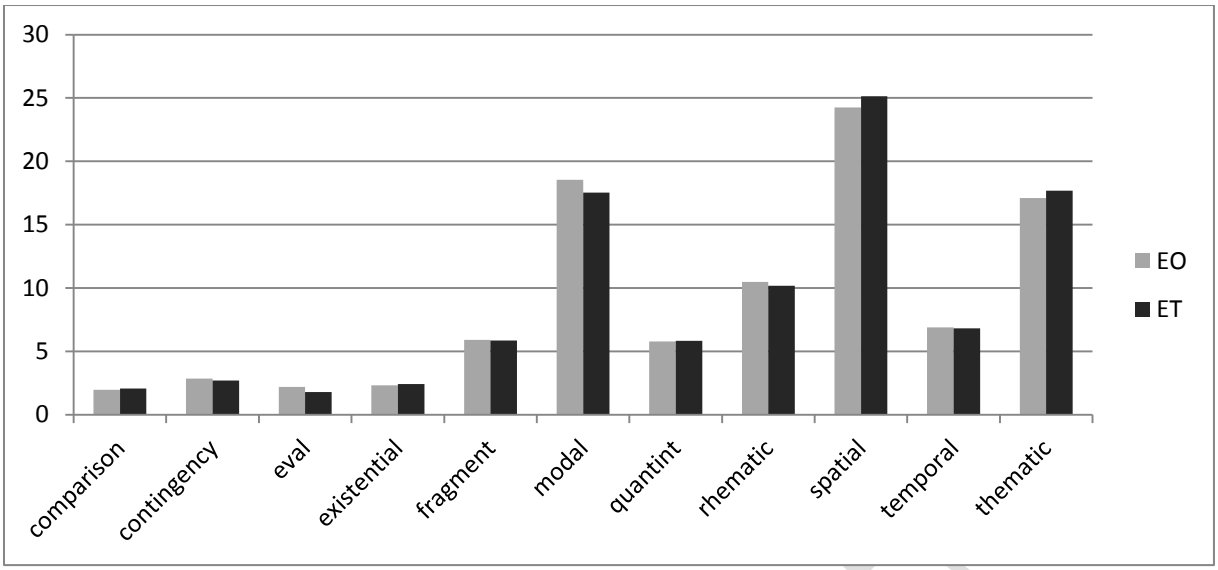

Figure 2. Adjusted proportional distribution of 3-gram tokens in EO and ET

Table 5. *p*-values calculated against two different baselines based on topped-up token counts: 3-grams and threshold tokens

Category	**tokens / 3-grams**	**tokens / threshold tokens**
Comparison	0.0024	0.0066
Contingency	0.6796	0.0856
Evaluative	0.2106	0.0140
Existential	0.6050	0.0747
Fragment	0.0011	0.1078
Modalizing	0.3133	<0.001
Organizational	0.0023	0.0282
Process	0.0164	0.0719
Quantifying/Intensifying	0.2992	0.9803
Reporting	0.0054	<0.001
Rhematic	0.5227	<0.001
Spatial	<0.001	<0.001
Temporal	<0.001	0.0023
Thematic stem	0.9913	0.0291

Table 6. Temporal 3-grams sorted by difference in number of tokens (raw frequencies)

Variety	Diff.	EO	Freq.	ET	Freq.
ET	167	a long time	128	a long time	295
ET	132	at the same	56	at the same	188
ET	117	the same time	57	the same time	174
EO	74	at the time	139	at the time	65
EO	69	by the time	120	by the time	51
EO	65	in the end	135	in the end	70
ET	52	now and then	51	now and then	103
ET	51	in the evening	24	in the evening	75
ET	50	and then i	42	and then i	92
ET	48	the whole time	10	the whole time	58
ET	47	and then she	24	and then she	71
ET	41	for the time	8	for the time	49
ET	38	the time being	6	the time being	44
ET	36	again and again	8	again and again	44
ET	36	for a while	152	for a while	188

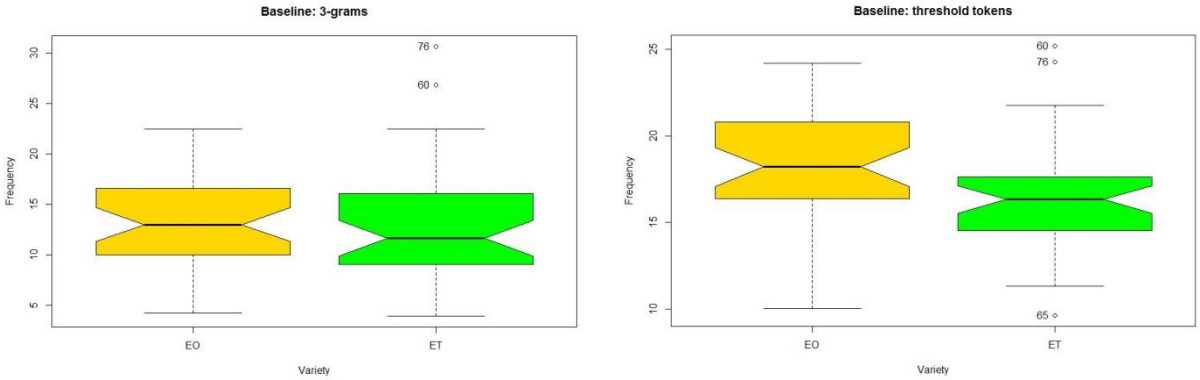

Figure 3. Boxplots of Thematic stem with two different baselines

Table 7. Ranking of normalized frequencies for Thematic stem based on two different baselines

Baseline: 3-grams		Baseline: threshold tokens	
EO	ET	EO	ET
2	1	2	1
3	4	3	4
5	7	8	5
6	10	9	6
8	11	11	7
9	12	12	10
15	13	14	13
18	14	23	15
21	16	25	16
22	17	29	17
...	...	...	...
61	59	65	46
62	60	66	47
65	63	67	52
66	64	68	54
67	68	70	56
69	70	72	58
71	72	73	69
73	76	74	71
74	77	75	77
75	78	76	78
1,590	1,491	1,796	1,285

Table A1. Comparing EO and BNC Baby on the basis of *p*-values

	Separate 3-gram type lists		Combined 3-gram type list	
Category	tokens / 3-grams	tokens / threshold tokens	tokens / 3-grams	tokens / threshold tokens
Comparison	0.5714	0.2711	0.8226	0.6926
Contingency	0.0456	<0.001	0.4856	0.1026
Evaluative	0.5203	0.2260	0.9636	0.7757
Existential	0.7759	0.4431	0.8333	0.5068
Fragment	<0.001	0.0023	0.1030	0.8189
Modalizing	0.5731	0.6459	0.4092	0.4549
Organizational	0.7394	0.2166	0.6778	0.2125
Process	0.9936	0.9454	0.9115	0.8236
Quantifying/Intensifying	0.1967	0.9569	0.2621	0.9562
Reporting	0.1486	0.0319	0.4451	0.1728
Rhematic	0.8568	0.2675	0.4197	0.9874
Spatial	0.0919	0.2918	0.3633	0.6977
Temporal	0.7046	0.1528	0.9004	0.2634
Thematic	0.6996	0.8745	0.6380	0.8841