Does corpus size matter? Revisiting ENPC case studies with an extended version of the corpus

Signe Oksefjell Ebeling, University of Oslo

Abstract
The validity of contrastive findings that base themselves on material from small parallel corpora may be questioned, and ever since the compilation of the English-Norwegian Parallel Corpus (ENPC) and English-Swedish Parallel Corpus (ESPC) some 20 years ago we have been aware of this. Recently, the ENPC has been expanded into the ENPC+, holding bidirectional translation data three times the size of the fiction part of the original ENPC. Drawing on material from the ENPC+, this paper replicates three contrastive studies made on the basis of the fiction part of the original ENPC to explore to what extent corpus size matters. The replica studies suggest that individual style, genre and date of publication are variables that may have a greater impact on the results than mere corpus size.

1. Introduction

Parallel corpora are generally small in size and the validity of contrastive findings based on these corpora may be questioned as a consequence. Size is here understood in terms of the number of tokens, or running words, making up the corpora. Being aware of the relatively moderate size of parallel corpora, including the English-Norwegian Parallel Corpus (ENPC) and English-Swedish Parallel Corpus (ESPC),[1] researchers have been on the cautious side when making use of these corpora. Indeed, warnings and comments such as the following have commonly been expressed ever since the compilation of these corpora in the mid-1990s:

- Due to its restricted size, the corpus is not suitable for studies of collocations and lexical studies beyond the core vocabulary. (S. Johansson 1998a: 11)
- […] occurrences are too few to allow any generalisations. (S. Johansson 2008: 111)

[1] ENPC: http://www.hf.uio.no/ilos/english/services/omc/enpc/; ESPC: http://sprak.gu.se/english/research/research-activities/corpus-linguistics/corpora-at-the-dll/espc

Ebeling, Signe Oksefjell. 2016. "Does corpus size matter? Revisiting ENPC case studies with an extended version of the corpus." *Nordic Journal of English Studies* 15(3):33-54.

- We need bigger corpora […] (S. Johansson 2009: 37)
- The number of examples is too small to be statistically reliable […] (M. Johansson 1996: 135)
- Provided that the material is large enough, MC values are thus a useful means of establishing semantic paradigms […] (Altenberg 1999: 266)
- Admittedly, the unit is not a very frequent one in the present data [...] (Ebeling et al. 2013: 191)

And finally, although mentioning the limited size of parallel corpora, Viberg (2010) has a more positive outlook on the matter:

- The limited size of this and several other parallel corpora is a temporary problem. (Viberg 2010: Section 2.1)

With regard to the first bullet point, it should be noted that contrastivists working with small-size parallel corpora have paid heed to size and have largely focused on high-frequency words, constructions or categories. This is in line with the commonly held view that "optimum corpus size depends on the specific linguistic investigation to be undertaken" (Granger 1998: 11, with reference to de Haan 1992).

The aim of this paper is to investigate to what extent the almost apologetic tone in the quotations above was justified when presenting some of the findings from the original parallel corpora, constantly drawing attention to the limited material at hand. With a bidirectional corpus three times the size of the fiction part of the original ENPC, viz. the ENPC+, the current paper will revisit a few case studies and compare findings based on the smaller (original) version of the corpus with those of the expanded version. The following lexis-based case studies will be considered:

- Ebeling (2003) on the Norwegian pseudo-coordination construction *bli* + present participle + *og* + infinitive;
- Johansson (1998b) on loving and hating in English and Norwegian;
- Johansson & Løken (1997) / Johansson (1998a) on some Norwegian discourse particles and their English correspondences.

The article is structured as follows: Section 2 begins with a description and comparison of the two versions of the ENPC. In Section 3, each of the original case studies is introduced in turn, and the findings are compared with those of the ENPC+ follow-up studies. Potential implications of the overall findings will be discussed in Section 4, while Section 5 offers some concluding remarks and future prospects.

2. The ENPC vs. the ENPC+

The structure and the contents of the ENPC have previously been described in several publications (e.g. Johansson & Hofland 1994, Johansson 1998a, Oksefjell 1999), and only a brief outline will be offered here. The corpus is parallel in the sense that it contains comparable fictional and non-fiction texts in English and Norwegian as well as translations of the texts from and into the two languages. This parallelism, and bidirectionality, is captured in the oft-repeated illustration in Figure 1.

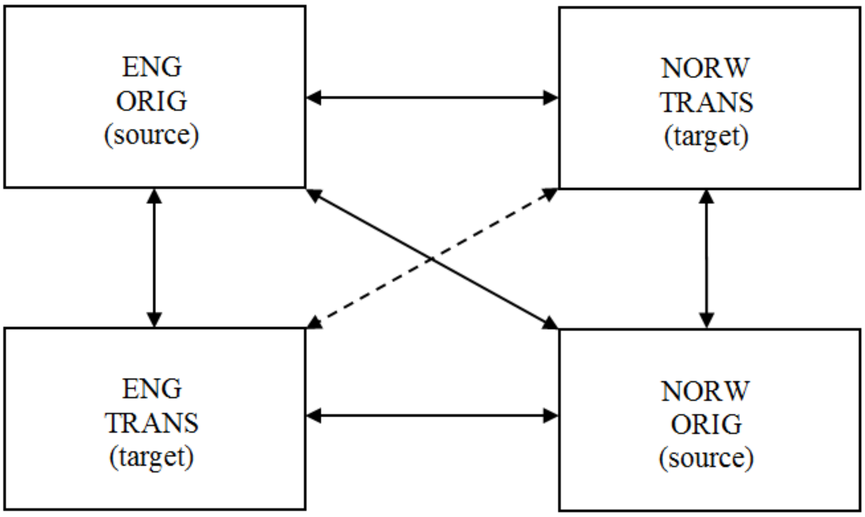

Figure 1. The structure of the ENPC (Johansson & Hofland 1994: 26)

It should be pointed out that the present investigation is concerned with the fiction part of the ENPC, as the expansion only includes fictional texts. Henceforth the original ENPC will be referred to as ENPCfiction.

The expansion of ENPCfiction, resulting in the ENPC+, was mainly prepared in 2011-2012, almost 20 years after the initial stages of the compilation of the ENPC (see Johansson & Hofland 1994). A full description of the ENPC+ and a comparison between ENPCfiction and its expansion are also offered elsewhere (Ebeling & Ebeling 2013: 86ff), but some of the main points are repeated here in Table 1.

Table 1. Main differences between ENPCfiction and its expansion (cf. Ebeling & Ebeling 2013: 87)

ENPCfiction	Expansion
Children's fiction, detective fiction and general fiction	Mainly crime (detective) fiction, some general fiction
Varieties of English (e.g. American, British, Canadian, South African)	Mainly British English
Texts (mainly) from the 1980s and 1990s	Texts from 2000 to 2012
Varieties of Norwegian (*bokmål* and *nynorsk*)	Norwegian *bokmål*
55 writers	13 writers
47 translators	10 translators
Extracts of 10,000–12,000 words	Complete books
400,000 x 2 words of original fiction material + their translations	900,000 x 2 words of original fiction material + their translations

With reference to Table 1, it is easy to point to variables that may have an impact on the comparison of studies based on ENPCfiction and the ENPC+, and it is not unproblematic to lump the two sets of data into one corpus. First of all, the fact that the expansion contains more text distributed across fewer writers and translators opens up for more idiosyncracies due to individual variation. Second, the texts contained in the expansion are naturally of a more recent date; thus diachronic change will have to be considered as a possible variable. Finally, variety and genre may also play a role, and perhaps genre in particular can skew the findings of the current investigation in certain directions. These potential pitfalls notwithstanding, the two parts have been brought together in the ENPC+.

An important issue in corpus linguistics in general, and another conceivable problem in the current study, is the degree of replicability of the studies. Although all the studies under scrutiny give relatively detailed information on how the actual linguistic classification was carried out, a 100% match between the original case study and the replica study cannot be guaranteed. Interestingly enough, this is also true of the study I performed myself some 12 years ago (Ebeling 2003), serving as a reminder of the importance of including precise descriptions of the criteria used in the classification of corpus data.

3. Studies revisited

The following sections focus on three contrastive topics previously studied on the basis of the original version of ENPCfiction. The studies chosen to be replicated are all well-known to me; one is self-authored and the others are, or have been, part of my taught syllabus.

Each original study will be introduced briefly before the ENPC+ material is added for comparison. First, Section 3.1 concerns itself with Ebeling's (2003) investigation of Norwegian *bli* 'remain' as part of pseudo-coordination and its English correspondences.[2] Section 3.2 takes a closer look at Johansson's (1998b) comparison of the English verbs *love*/*hate* and their Norwegian counterparts *elske*/*hate*. Finally, Section 3.3 deals with Norwegian discourse particles and their English correspondences, using Johansson & Løken (1997) as a starting point, but mainly drawing on Johansson (1998a), who shifts the focus of attention to *probably*, one of the English correspondences of the Norwegian discourse particle *nok*.

3.1 Ebeling (2003)

The Norwegian sequence *bli* 'remain' + present participle + *og* 'and' + infinitive is a type of pseudo-coordination, i.e. it is a construction expressing hypotaxis rather than parataxis, even if the coordinating conjunction *og* is present (Vannebo 1969). Pseudo-coordination in Norwegian mainly consists of a posture verb + coordinated verb, as shown in example (1). In addition, the combination dealt with here has *bli* 'remain' as an auxiliary verb.

(1) Han <u>blir stående og se</u> seg om, forundret, for alle er helt stille, og det er ikke Pappen akkurat vant til. [LSC1]
Lit.: He remains standing and look around …

[2] Note that *bli* is a highly polysemous and versatile verb corresponding to a variety of English high-frequency verbs – notably *be*, *become*, *get* and *remain* – depending on context (see further Ebeling 2003).

This study was part of a larger contrastive investigation of the two Norwegian verbs *bli* and *få* and their correspondences in English; my focus in the sub-study of *bli* as part of pseudo-coordination was inevitably on what happened to *bli* in translation between English and Norwegian.

The distribution of types of English translations is given in Table 2, where it is shown that a so-called "synthetic" translation is found in almost 80% of the cases, as shown in the column in greyscale (ENPCfiction). By synthetic is meant cases where *bli* + present participle are merged into one verb, i.e. the one corresponding to the present participle, as in examples (2) and (3).

(2) Han blir stående og se seg om, forundret, for alle er helt stille, og det er ikke Pappen akkurat vant til. [LSC1]
He stands there, looking around, surprised because everyone is completely still, and Woody is not exactly used to that. [LSC1T]

(3) Jeg blir stående og lytte. [ToEg1N]
I stand listening to it for a while. [ToEg1TE]

Table 2. English translations of *bli* when *bli* is followed by a present participle + *og* + infinitive in ENPCfiction vs. ENPC+[3]

	ENPCfiction	ENPC+
translation	no.	no.
keep	1	5
be	2	4
remain	4	8
Ø	4	19
'synthetic'	43 (79.6%)	114 (76%)
Total	54	150

As shown in Table 2, other, marginal, translation types include *keep*, *be*, *remain* and Ø. An example of *remain* as a translation of *bli* is given in (4).

(4) Midt på gulvet ble hun stående og se seg om en god stund, og jeg ventet engstelig. [EHA1]

[3] The ENPCfiction part of Table 2 is taken from Ebeling (2003: 169).

> In the center of the room she remained standing, looking around for a few moments while I waited anxiously. [EHA1T]

54 occurrences of pseudo-coordination with *bli* were recorded in ENPCfiction. In the ENPC+ this is almost trebled to 150 occurrences, as shown in Table 2. Given that the ENPC+ is three times the size of ENPCfiction, this corresponds to the expected increase. In addition, a fairly similar distribution of translation types is found in the ENPC+ study. Thus, the relationship between the Norwegian construction and its English translation patters appears to be stable.

The picture emerging from this overview suggests that, in the case of this particular construction, similar conclusions can be drawn on the basis of less data. One of the conclusions of the original study at this stage of the discussion was that there is no clear English counterpart of *bli* as part of this construction, although a few examples do occur, as with *remain* in example (4) above.

The main finding, however, was that the auxiliary *bli* tends to be absorbed in the posture verb in translations into English, which tries to capture the continuative nature of Norwegian *bli* + present participle. This is particularly evident in instances such as (2), where the adverb *there* has been added in the translation, seemingly to get the durative element of the Norwegian construction more clearly across in the English translation (see also Ebeling 2015b). A similar conclusion can be drawn on the basis of the ENPC+ material; thus, size does not seem to matter in this case.

3.2 Johansson (1998b)[4]

Before I introduce the original study by Johansson, it should be mentioned that this particular study has also been revisited in a couple of other papers, albeit with a different focus: Hasselgård (2011) in a contrastive study of spoken English and Norwegian and Ebeling (2015a) in a contrastive study of written English and Portuguese. Both of these offer interesting similarities with, and additions to, the present

[4] A slightly revised version of the article is published as Chapter 5 in Johansson (2007).

investigation, providing broader insights into the discussion of love and hate verbs across languages.

Johansson's study entitled "Loving and hating in English and Norwegian" was triggered by a couple of sentences appearing in Norwegian newspapers; one of which is repeated here as example (5). Johansson established that this news item was a direct translation from English, and the original version is offered below the Norwegian example (cf. Johansson 1998b: 93).

(5) Jeg hater å bringe sladderen videre.
(The original version: 'I hate to pass gossip on' was attributed to Shirley MacLaine and quoted in *Østlandets Blad*)

Johansson's immediate reaction was that the use of Norwegian *hate* in this context did not ring quite idiomatic, and he decided to examine the relationship between English *hate* and Norwegian *hate*, and also added their more loveable opposites: *love* and *elske*. More specifically, he investigates the relationship between these verbs in English and Norwegian, with particular attention to the types of objects they typically occur with. He applies the broad categories of *personal* and *non-personal* object. The distribution, referred to as percentages in the original study, is found in Table 3.[5]

Table 3. Distribution of objects with *hate, elske, hate* and *love* (the ENPCfiction distribution from Johansson 1998b: 95)

	ENPCfiction		ENPC+	
	Personal objects	Non-personal objects	Personal objects	Non-personal objects
N hate	65%	35%	54%	46%
N elske	61%	39%	59%	41%
E hate	27%	73%	19%	81%
E love	46%	54%	49%	51%

Johansson notes that in Norwegian, "the verbs take a personal object in the majority of cases, while non-personal objects are more common in English original texts" (p. 95). Examples include (6), where Norwegian

[5] Johansson also looks into the use of these verbs in the translated texts and notes some discrepancies between originals and translations. This study, however, will limit itself to the original texts only.

elske is followed by a personal object, and (7), where English *hate* is followed by a non-personal object in the form of a non-finite clause.

(6) Jeg trodde henne når hun stadig fortalte at hun elsket meg. [JW1]
I believed her when she constantly told me that she loved me. [JW1T]

(7) He did once tell me that he hated shaking hands. [RDA1]

As can be seen in the ENPC+ columns in Table 3, this is still the case, but with some notable differences with regard to proportion. While the numbers for *elske* and *love* are more or less unchanged, the situation for *hate* and *hate* is different, as there is a marked increase of non-personal objects in both languages in going from ENPCfiction to the ENPC+. More than anything, and as the overall frequency of these verbs does not deviate much from the overall expected increase (i.e. the frequency is expected to treble, see Table 4), this seems to point to a change in use of the two verbs, rather than being a matter of having more data at our disposal.

Table 4. Comparison of total distribution of the verbs in ENPCfiction vs. ENPC+ (the ENPCfiction figures are taken from Johansson 1998b: 95)

	ENPCfiction	ENPC+
	Original texts	Original texts
N hate[6]	23	59
N elske[7]	36	122
E hate	67	212
E love	100	303

This potential language change is substantiated by the fact that Johansson found no occurrences in his material of *hate* + clausal complement in the Norwegian original texts, while there, in the ENPC+ material, are three such instances, one of which is found in example (8).

[6] The form *hata* 'hated' seems to have been excluded from Johansson's study; thus, this form has also been left out here (six occurrences in the material altogether).

[7] The form *elska* 'love/loved' seems to have been excluded from Johansson's study; thus, this form has also been left out here (seven occurrences in the material altogether).

(8) Men hun hatet å jogge. [JoNe2N]
But she hated jogging. [JoNe2TE]

This is precisely the kind of complementation that triggered the original study; Johansson states that his "immediate reaction was that these were anglicisms" (p. 93). My guess is that he was right; moreover, the more recent data also suggest that this pattern is on the increase in Norwegian. Johansson notes that, "[c]hanges of this kind are natural wherever there are languages in contact, but it is important to be aware of what is going on" (p. 102). The fact that such a change has taken place, or is taking place as we speak, was also confirmed by Hasselgård (2011) who performed an English-Norwegian contrastive study of spoken material of a more recent date than the written material in ENPCfiction.

Furthermore, it seems to have become more common, both in English and Norwegian to hate non-personal objects, thus suggesting that the force of the verbs may have been weakened. As pointed out by Johansson (p. 101):

> Whereas Norwegian *hate* and *elske* express a strong feeling and typically with a personal object, English *hate* and *love* are also used in a weakened sense […]. The weakened sense is most likely to appear where the verbs combine with non-personal objects, particularly complement clauses.

This tendency of achieving a more weakened sense seems to be on the increase in both languages as far as *hate* is concerned; however, the three instances of complement clauses following N *hate* notwithstanding, the increase seems to be more prominent in non-personal NP complementation, as the percentage of complement clauses seems to be fairly stable (in Johansson's material, 26.5% of the non-personal objects of E *hate* were complement clauses, while the percentage in the ENPC+ material is 27.2%).[8]

Johansson himself draws attention to the question of corpus size, and says: "Judging by our limited material, it seems as if *elske* is more compatible with a following infinitive than *hate*" (Johansson 1998b: 99). This seems only to be marginally the case in the ENPC+ material, where 12% of the non-personal objects of *hate* and 13% of the non-personal

[8] In the ENPC 13 out of 49 non-personal objects were complement clauses; in the ENPC+ 34 out of 125 were complement clauses.

objects of *elske* are infinitive clauses. Thus, I will maintain that, rather than increased size, it is the more recent text material of the ENPC+ that explains the difference in non-personal objects with *hate*/*hate*; Norwegian in particular seems to be undergoing a change in accepting these complementation patterns more readily than was the case 20-30 years ago.

3.3 Johansson & Løken (1997) / Johansson (1998a)

Pre-dating the completion of the ENPC, these two studies were based on a slightly smaller sample than ENPCfiction. More specifically, the sample contains 27 fiction texts in each direction of translation instead of 30 (approx. 360,000 words in each part of the corpus, amounting to roughly 1.4 million words in total compared to the ENPC+ with 5.2 million; i.e. the ENPC+ is 3.7 times larger than the version of the ENPC used in the two studies. This will of course be taken into consideration in the comparison below.

One of the discourse particles Johansson & Løken deal with in some detail is *nok* and its correspondences in English. Table 5 gives an overview of the distribution of these correspondences – both translations and sources – in the ENPCfiction material.

Table 5. Correspondences of the Norwegian modal particle *nok*, expressed in percent within each column (cf. Johansson & Løken 1997: 168-169; Johansson 1998a: 14)

	ENPCfiction	
Correspondence	E translations (N = 141)	E source (N = 79)
probably	25	6
other adverb	21	4
verb construction	11	10
clause	9	10
miscellaneous	3	5
zero	31	65

As mentioned in Section 3, the study by Johansson & Løken (1997) on Norwegian discourse particles and their English correspondences served as a starting point for Johansson (1998a). While the discussion of *nok* in Johansson & Løken (1997) is concerned with all of its English correspondences and the lack of a clear English counterpart, as evidenced by the strikingly high number of zero correspondences, Johansson (1998a) continues the discussion by looking at it from the reverse perspective, focusing on the main single overt translation correspondence *probably*.

He starts by mapping the Norwegian correspondences of *probably*, and notes that the frequency of zero correspondences is low and Norwegian modal particles are infrequent as translations of *probably* (see Table 6). "A plausible interpretation of these results", he says, "is that the existence of close formal and semantic correspondences simplifies the task of the Norwegian translator. By contrast, when faced with the problems of rendering Norwegian *nok*, the English translator finds no easy solution" (p. 15).

Table 6. Correspondences of the English adverb *probably*, expressed in per cent within each column (cf. Johansson 1998a: 15)

	ENPCfiction	
Correspondence	N translation (N = 94)	N source (N = 141)
nok	3	25
vel	6	28
antagelig(vis) [antakelig]	21	3
kanskje	3	9
sannsynligvis	37	16
sikkert	11	9
trolig	3	1
miscellaneous	13	6
zero	2	4

The concern in the present investigation is whether the same conclusions can be drawn when the study is based on material from a more sizeable corpus. The replica study will focus on *probably* and its translations into Norwegian only.

Comparing the number of occurrences in the two versions of the corpus (see Table 7), we can observe a striking difference in the frequency with which *probably* occurs: 94 vs. 580, which means that it is more than six times as frequent in the ENPC+ compared to ENPCfiction. Given that the ENPC+ is 3.7 times larger than the version of the ENPC that Johansson used, the expected frequency of *probably* in the ENPC+ would be around 350. Another striking observation that can be made from Table 7 is the use of *antagelig, antakelig* and *antageligvis,* for which the most frequent variant – *antakelig* – is used as shorthand for all three, in the more recent material. The *antakelig*-words in Johansson's material accounted for 21% of the occurrences, while they account for almost 50% in the ENPC+ material. This increase seems to have taken place at the expense of *sannsynligvis* in particular (but also other items in the miscellaneous category).

Table 7. Translations of the English adverb *probably*, expressed in per cent within each column, in ENPCfiction vs. ENPC+

	ENPCfiction	ENPC+
Correspondence	N translations (N = 94)	N translations (N = 580)
nok	3	5.1
vel	6	7.4
antakelig	21	47.7
kanskje	3	1.6
sannsynligvis	37	10.7
sikkert	11	16.5
trolig	3	1.9
miscellaneous	13	6.4
zero	2	2.7

In the study on *love* and *hate* (Section 3.2), linguistic change as a result of the time span between the two parts of the corpus seemed to account for some of the differences that were noted between the two versions of the corpus. In the current context, however, it is less likely that the increased use of *probably* is due to linguistic change. In fact, there rather seem to be two writers (and three texts) in particular that contribute towards the increased use of *probably* in the ENPC+, namely PeRo1E (90 occ.), PeRo2E (102 occ.) and TaFr1E (114 occ.).[9]

Admittedly, the word count for these texts is around ten times higher than the word count for the text extracts of ENPCfiction; nevertheless, the use of *probably* is proportionally much higher in those three texts.

In some sense, the three texts referred to can be characterised as what Sinclair calls "rogue texts":

> In any variety of a language there will be some texts — "rogue" texts — which stand out as radically different from the others in their putative category. (Sinclair 2005: 13)

[9] PeRo = Peter Robinson; TaFr = Tana French. See Ebeling & Ebeling (2013: 241ff) for a full overview of authors and texts included in the ENPC+.

While the texts may not be so different from the others in their category overall, they are radically different in the use of particular words. In the light of Sinclair's observation regarding "rogue texts", an experiment of excluding the three texts in question was conducted. Thus a version of the ENPC+ 2.3 times larger than the version of ENPCfiction used by Johansson was produced.

When excluding the three texts in question, it can be seen, in Table 8, that the number of occurrences of *probably* increases by 2.9 compared to ENPCfiction (instead of 6.2, which was the case above), which is much closer to the expected increase of 2.3. A log-likelihood test shows that the difference is not statistically significant.

Table 8. Number of occurrences of *probably* in the reduced version of the ENPC+ vs. ENPCfiction

ENPCfiction	ENPC+ (excluding PeRo1E, PeRo2E, TaFr1E = 835,000 words = 2.3 times the size of Johansson's original sample	
(N = 94)	(N =274)	LL = 3.79; p > 0.05 (difference is not statistically significant)

However, while *probably* is now seen to increase by roughly what would be expected, the translation correspondences still show a bias in the use of *antakelig* at the expense of *sannsynligvis*, Although one translator in particular seems to favour *antakelig*, this does not fully explain what is going on – whether *antakelig* is generally on the increase at the expense of *sannsynligvis*, or whether it is only tied to the preferences of the individual writers and translators represented in the corpus.

Such observations are of course far from new: corpus linguists have always been aware of individual variation and idiosyncracies on the part of the writers (and translators), and it seems to be a more pertinent problem in the use of individual words than in the use of grammatical patterns. Individual lexical items appear to stand a greater risk of turning into some sort of pet word. In the future, proper statistical measures on these issues should be carried out, but it is beyond the scope of this paper.

In connection with the use of words such as *probably* and *antakelig* it is also important to mention that they are often used in dialogue, and dialogue is a feature that, according to de Haan (1996), is more characteristic of crime fiction than general fiction.

> It is obvious that the three crime texts contain far more dialogue than any of the other four texts (roughly 60 per cent of the total number of sentences in each of the three crime texts). (de Haan 1996: 26)

The three texts that were excluded are all in the category "crime fiction", which means that they are expected to contain more lexical items typical of dialogue, thus chances are that the individual writers and translators will use more dialogue-prone items.

On a more general note regarding dialogue in fiction, Axelsson (2009: 191) draws attention to the fact that "[o]ne cannot take it for granted […] that the amount of direct speech or the linguistic features of direct speech are similar in samples from different parts of the fiction texts (beginning, middle and end samples)." In the present context, however, the samples are considered homogeneous and therefore comparable, as all the texts in the original ENPCfiction are taken from the beginning of texts, while the texts making up the extension are all full texts.

Axelsson (ibid.: 191-192) also mentions the lack of mark-up of direct speech in existing corpora as a problem. Although such mark-up is in fact available in the original ENPCfiction, a systematic discussion of items typically found in dialogue vs. narrative was not considered essential to the current analysis.

The results of this study seem to paint a more complex picture than the other revisited case studies regarding the impact of corpus size in contrastive studies. Although the size of the corpora definitely plays a role in that a formidable increase in the number of attested instances of *probably* and *antakelig* can be noted, the increase may also be attributed to individual preferences and genre. In other words, there is no evidence that corpus size alone accounts for the differences in distribution of these items in ENPCfiction vs. the ENPC+, suggesting that the original observations made on the basis of the smaller ENPCfiction may still be valid for the two languages in question.

4. Corpus size in contrastive studies: Summing up

In order to discuss the implications of corpus size in contrastive studies, it is important to return to the underlying questions and concerns that triggered the present study: Does corpus size matter?; and is the original ENPC large enough to yield valid results?

The experiment of replicating previous ENPC studies on the basis of a bigger corpus has thrown up various, and perhaps not new and unexpected, issues related to corpus size and the nature of corpora in general. However, corpus size alone may have less impact than was perhaps feared. Given the design and structure of the original ENPC, where care was taken to include as many different writers and translators as possible, we can rely on the findings even if the material feels limited in terms of number of tokens. Nevertheless, it should be noted that bigger parallel corpora are of course needed for studies of less frequent items and constructions, as well as for studies of recurrence. Moreover, as noted by Johansson (2011: 128):

> Although there are many advantages with the bidirectional translation model, corpora built in this way may need to be supplemented by larger corpora compiled according to the two main models presented above [comparable monolingual corpora and unidirectional translation corpora], as these are less constrained with respect to the types and range of texts.

It has also been confirmed that the issue of the individual is far from a trivial one; the need to keep an eye on dispersion and potential rogue texts is certainly upheld, particularly, it seems, in studies of lexis. The contrastive nature of the studies reported on here is of course a complicating factor, as it involves two versions of the same text, viz. the original and its translation, each potentially marked by the author's or translator's individual style. And, not unexpectedly, it seems as if discourse particles and adverbs of the kind discussed here are more easily subject to individual preferences both on the part of the writers and the translators, and are thus more likely to skew the results. In addition, lexical choices may also be influenced by genre, as in the case of items typical of dialogue in crime fiction. In this respect individual style and genre seem to be variables that have a greater impact on the results than mere corpus size.

Furthermore, we have seen that it is not unproblematic to expand parallel corpora in the way done for the ENPC+, particularly because of the time lag between the texts in the original ENPC and those making up the expansion. This was seen as a particularly decisive factor in the love-hate study in Section 3.2.

The investigation started out by asking whether corpus size matters, and the first previous ENPC study that was probed into, on the English correspondences of *bli* + past participle (Section 3.1), showed that it does not necessarily matter. In other words, corpus size had little effect on the findings and conclusions similar to those drawn in the original study could also be drawn on the basis of the larger data set.

However, the studies revisited in Sections 3.2.and 3.3 resulted in a more complex answer to the initial question, involving, it could be claimed, different kinds of rogueness. Figure 2 is an attempt at capturing (some of) the multi-faceted nature of such rogueness.

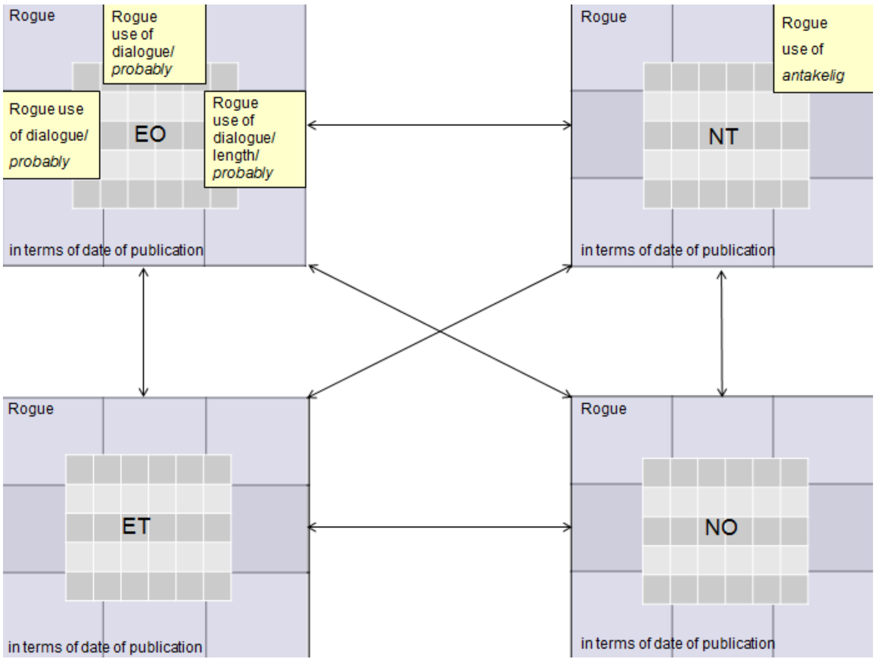
Rogue
Rogue use of dialogue/ *probably*
Rogue use of dialogue/ *probably*
EO
Rogue use of dialogue/ length/ *probably*
in terms of date of publication
Rogue
Rogue use of *antakelig*
NT
in terms of date of publication
Rogue
ET
in terms of date of publication
Rogue
NO
in terms of date of publication

Figure 2. Some types of rogueness identified in the ENPC/ENPC+, including date of publication, use of specific lexical items, use of dialogue and length of text

In the second study (Section 3.2), the findings on the use of *love* and *elske* were found not to be affected by corpus size or any kind of rogueness, while the findings for *hate* and *hate* suggested that a language change has taken place or, indeed, is ongoing; thus, the more recent texts can be described as rogue in terms of publication date. This is indicated in Figure 2 by the area surrounding the 30x4 squares illustrating the original ENPCfiction; all of the nine full texts constituting the extension – represented by the larger squares in Figure 2 – are rogue in terms of publication date.

The final study (Section 3.3) was the one that really drew our attention to rogueness and the role of the individual, i.e. the writer and the translator. Rogueness in the use of specific lexical items was identified, as well as rogueness in the use of dialogue in three crime fiction texts in particular; additionally one of those texts is longer than the other texts in the corpus. These issues are illustrated (in Figure 2) in the squares representing three texts in English originals (EO) and one square representing one text in Norwegian translations (NT).

With such a broad definition of rogueness as the one adopted here, a text can be rogue in one area of study (*probably*) but mainstream in another area of study (*bli* + present participle). This calls for caution on the part of (parallel) corpus compilers in the future; it will be all the more important to include a wide variety of writers and translators and to check for rogueness in each study that is carried out.

5. Conclusion and future prospects

Even if this study has highlighted some problematic issues related to corpus size and structure, these can to a large extent be controlled for, and contrastive studies based on existing parallel corpora still yield valid and sound results. However, as hinted at in Section 3.3, corpus-based contrastive studies would in the future benefit from more sophisticated statistical treatments for analysing the impact of the different variables involved. One potential variable that may play a role when it comes to linguistic choices is language variety, e.g. British vs. American English.[10] This was mentioned when outlining the differences between ENPCfiction and its extension but was not discussed in connection with the studies revisited. It is, nevertheless, a factor worth taking into consideration, particularly if robust statistical tests were implemented.

Moreover, Axelsson (2011), in her cross-linguistic study of tag questions, points to another restriction on the data culled, in her case, from the ESPC, namely the fact that only samples from the beginning of

[10] See also Axelsson (2011: 219), who draws attention to this as a factor potentially contributing to certain preferred usage patterns.

texts are included. Thus, following Sinclair (2005: 6), it could be argued that "ideally, documents and transcripts of verbal encounters should be included in their entirety". However, copyright holders rarely donate complete texts to a corpus, thus opening up for including different parts of texts, e.g. beginning, middle and concluding samples. This would, unfortunately, introduce yet another variable that may or may not prove significant in the study of language use.

With a pragmatic approach to what is feasible to obtain, Johansson seems to have struck the right balance for devising an appropriate structure for bidirectional parallel corpora (see Johansson & Hofland 1994; Johansson et al. 1999/2002). As a future prospect, it should therefore be encouraged to compile parallel corpora matching the existing ones in terms of content and structure, but comprising texts of a more recent date. In a similar fashion to what has been done for the LOB and Brown corpora – with FLOB and Frown[11] – a carefully designed ENPC 20 years on would pave the way for a new field of diachronic corpus-based contrastive studies, ensuring that such studies can be carried out in a systematic way.

References

Altenberg, Bengt. 1999. "Adverbial connectors in English and Swedish: Semantic and lexical correspondences." *Out of Corpora: Studies in Honour of Stig Johansson*. Eds. Hilde Hasselgård & Signe Oksefjell. Amsterdam: Rodopi. 249–268.

Axelsson, Karin. 2009. "Research on fiction dialogue: Problems and possible solutions." *Corpora, Pragmatics and Discourse*. Eds. Andreas H. Jucker, Daniel Schreier & Marianne Hundt. Amsterdam: Rodopi. 189–201.

Axelsson, Karin. 2011. *Tag Questions in Fiction Dialogue*. Ph.D. thesis. Göteborg: University of Gothenburg. [Available online at http://hdl.handle.net/2077/24047]

[11] The Lancaster-Oslo/Bergen (LOB) and Brown corpora are carefully designed 1-million-word monolingual corpora of British and American English, respectively, containing texts from 1961. Their 1991 counterparts – FLOB and Frown – were compiled according to the same criteria at Freiburg University, hence the 'F', to enable diachronic studies of British and American English (see the ICAME corpus collection at http://clu.uni.no/icame/newcd.htm).

de Haan, Pieter. 1992. "The optimum corpus sample size?" *New Directions in English Language Corpora*. Ed. Gerhard Leitner. Berlin / New York: Mouton de Gruyter. 3–19.

de Haan, Pieter. 1996. "More on the language of dialogue in fiction." *ICAME Journal* No. 20, 23–40.

Ebeling, Jarle & Signe Oksefjell Ebeling. 2013. *Patterns in Contrast*. Amsterdam: John Benjamins.

Ebeling, Jarle, Signe Oksefjell Ebeling, & Hilde Hasselgård. 2013. "Using recurrent word-combinations to explore cross-linguistic differences." *Advances in Corpus-based Contrastive Linguistics. Studies in Honour of Stig Johansson*. Eds. Karin Aijmer & Bengt Altenberg. Amsterdam: John Benjamins. 177–200.

Ebeling, Signe Oksefjell. 2003. *The Norwegian Verbs bli and få and their Correspondences in English. A Corpus-based Contrastive Study*. Oslo: Acta Humaniora.

Ebeling, Signe Oksefjell. 2015a. "Loving and hating in English and Portuguese: A corpus-based contrastive study." *Linguística, Informática e Tradução: Mundos que se Cruzam. Homenagem a Belinda Maia*. Eds. Alberto Simões, Anabela Barreiro, Diana Santos, Rui Sousa-Silva & Stella E. O. Tagnin. *Oslo Studies in Language* (OSLa) 7(1): 439–456.

Ebeling, Signe Oksefjell. 2015b. "A contrastive study of Norwegian pseudo-coordination and two English posture-verb constructions". *Cross-linguistic Perspectives on Verb Constructions*. Eds. Signe Oksefjell Ebeling & Hilde Hasselgård. Newcastle upon Tyne: Cambridge Scholars. 29–57.

Granger, Sylviane. 1998. "The computer learner corpus: a versatile new source of data for SLA research." *Learner English on Computer*. Ed. Sylviane Granger. London / New York: Longman. 3–18.

Hasselgård, Hilde. 2011. "Loving and hating in English and Norwegian speech." Paper presented at the Jan Svartvik Birthday Symposium, Lund, 19 August 2011.

Johansson, Mats. 1996. "Contrastive data as a resource in the study of English clefts." *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies, Lund 4-5 March 1994*. Eds. Karin Aijmer, Bengt Altenberg & Mats Johansson. Lund: Lund University Press. 127–152.

Johansson, Stig. 1998a. "On the role of corpora in cross-linguistic research." *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Eds. Stig Johansson & Signe Oksefjell. Amsterdam: Rodopi. 3–24.

Johansson, Stig. 1998b. "Loving and hating in English and Norwegian: A corpus-based contrastive study." *Perspectives on Foreign and Second Language Pedagogy*. Eds. Dorte Albrechtsen, Birgit Henriksen, Inger M. Mees & Erik Poulsen. Odense: Odense University Press. 93–103.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies.* Amsterdam/Philadelphia: John Benjamins.

Johansson, Stig. 2008. *Contrastive Analysis and Learner Language: A Corpus-based Approach*. Compendium for the course ENG2152 – Contrastive and Learner Language Analysis, University of Oslo.

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

Johansson, Stig. 2011. "A multilingual outlook of corpora studies." *Perspectives on Corpus Linguistics*. Eds. Vander Viana, Sonia Zyngier & Geoff Barnbrook. Amsterdam/Philadelphia: John Benjamins Publishing Company. 115-129.

Johansson, Stig, Jarle Ebeling & Signe Oksefjell. 1999/2002. "English-Norwegian Parallel Corpus: Manual." www.hf.uio.no/ilos/english/services/omc/enpc/ENPCmanual.pdf

Johansson, Stig & Knut Hofland. 1994. "Towards an English-Norwegian Parallel Corpus". *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zurich 1993*. Eds. Udo Fries, Gunnel Tottie, & Peter Schneider. Amsterdam: Rodopi. 25–37.

Johansson, Stig & Berit Løken. 1997. "Some Norwegian discourse particles and their English correspondences." *Sounds, Structures and Senses. Essays Presented to Niels Davidsen-Nielsen on the Occasion of his Sixtieth Birthday*. Eds. Carl Bache & Alex Klinge. Odense: Odense University Press. 149–170.

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

Sinclair, John. 2005. "Corpus and text – Basic principles." *Developing Linguistic Corpora: A Guide to Good Practice*. Ed. Martin Wynne. Oxford: Oxbow Books. 1–16. [Available online at http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm]

Viberg, Åke. 2010. "Basic verbs of possession: A contrastive and typological study." *ConTextes* Vol. 4 – *Unison in Multiplicity: Cognitive and Typological Perspectives on Grammar and Lexis*. http://cognitextes.revues.org/308