Parallel corpora and contrastive studies[1]

Hilde Hasselgård

University of Oslo

Abstract: For a long time corpus studies meant monolingual studies. Multilingual and parallel corpora have been available only since the late 1990s. The first machine-readable parallel corpora were the English-Norwegian Parallel Corpus and its sister project the English-Swedish Parallel Corpus. Just like monolingual corpora have led to new insights and new practices in descriptions of individual languages, parallel corpora have opened up new avenues of contrastive studies. Being machine-readable they can give faster access to more material than was previously possible on the basis of non-electronic parallel texts. They also make it easier to see cross-linguistic patterns of correspondence. The present paper touches on the development and use of multilingual corpora with a focus on work done in Scandinavia with the purpose of showing how parallel corpora can be useful within different fields of language description: lexis, grammar and discourse. It also presents a case study that demonstrates how a parallel corpus can be used in comparing two seemingly equivalent future-referring expressions cross-linguistically, namely the English 'be going to' and the Norwegian 'kommer til å' ('come to').

1 From monolingual to multilingual corpus linguistics

Corpus linguistics can be described as a methodology for studying language by means of (relatively) large, structured databases of text compiled and prepared for use in linguistic research (see Leech 2010: 104). Early developments of corpora and corpus linguistics methodology took place largely within English linguistics, with the Brown corpus as the first machine-readable corpus (1960s), followed by the Lancaster-Oslo/Bergen (LOB) corpus (see e.g. Svartvik 1992: 8 ff). The availability of corpora greatly facilitated the access to large amounts of linguistic material and opened up new possibilities for quantitative studies and variation studies. The fact that the first two widely used corpora (Brown and LOB) were compiled according to the same design criteria encouraged comparative studies of the two varieties of English; in other words, a comparative perspective was by no means foreign to corpus methodology. However, for a long time, corpora remained monolingual, and to the extent that contrastive studies could be corpus-based at all, they would have had to rely on monolingual corpora in different languages.

Multilingual corpora are a more recent development.[2] It was not until the early 1990s that Stig Johansson and Knut Hofland launched plans for the English-Norwegian Parallel Corpus (ENPC); see Johansson & Hofland (1994). The projected corpus would contain original texts in both English and Norwegian with translations into the other language. This type of corpus would require new technology for the alignment of originals and translations as well as data retrieval. Like other projects initiated by the late Stig Johansson, the English-Norwegian Parallel Corpus was realised, in close cooperation with Swedish colleagues who compiled the English-Swedish Parallel Corpus (ESPC) using the same design criteria and partly the same English original texts (Aijmer & Altenberg 1996: 79 ff.).

Software for alignment was developed by Knut Hofland (Hofland 1998) while Jarle Ebeling developed a system for parallel corpus concordancing (Ebeling 1998). Originals and translations in the ENPC are aligned at sentence level. Each sentence (or s-unit) has a unique identification tag with a pointer to the corresponding s-unit in the other language, so that e.g. a search in English originals for ‘language’ will bring out all the s-units containing this word along with their linked-up translations into Norwegian. An example is given in (1).

(1a) <s id=ABR1.1.1.s326 corresp=ABR1T.1.1.s325>But how come you speak the language so fluently?"</s> (ABR1)

(1b) <s id=ABR1T.1.1.s325 corresp=ABR1.1.1.s326>Men hvordan har det seg at De snakker språket så flytende?"</s> (ABR1T)

The identification tag of (1a) shows which text the example is from, in this case one by Anita Brookner, the number of the s-unit counting from the start of the sample, and a pointer to the corresponding s-unit in Norwegian. Note that the s-unit number of (1a) is 326 while that of (1b) is 325. This is because sentence boundaries are not always carried over from the original to the translation. The tag in round brackets at the end of (1b) ends in T, which shows that the example is a translation.

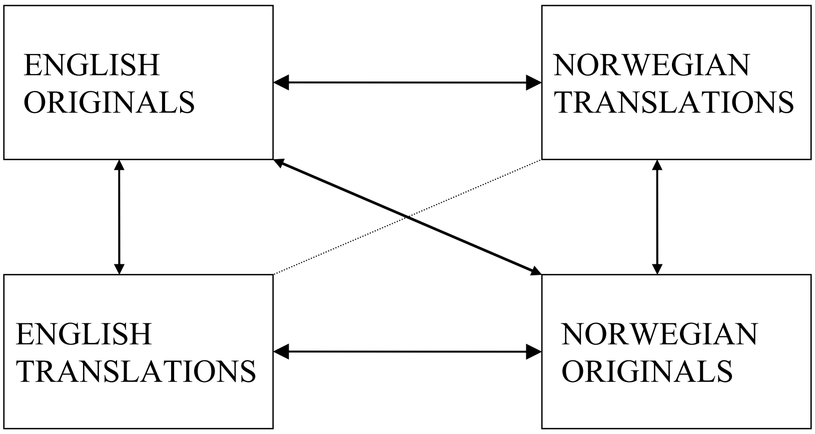

Figure 1. The structure of the English-Norwegian Parallel Corpus

The ENPC and the ESPC were designed according to the model shown in Figure 1: they contain original texts in both languages, distributed over fiction and non-fiction.[3] Translations into the other language are published translations by professional translators, i.e. the translations as well as the originals are authentic in the sense that they were not carried out for the purpose of being included in a corpus. The model can be characterized as a *bidirectional* *translation corpus*. Its design makes it possible to carry out a number of comparisons, as indicated by the arrows in the figure: (1) originals and their translations; (2) original texts of the same type in both languages; (3) translations in both languages; (4) original and translated texts in the same language. Thus the corpus can be used for contrastive studies (particularly using the first two types of comparison) and translation studies (types 1, 3 and 4).

It must be mentioned at this point that the ENPC model is not the only type of corpus that has been termed a parallel corpus. The term is often used about a corpus of original texts with translations into one or more other languages, i.e. a unidirectional *translation corpus*. Less commonly, the term is also used about a corpus of comparable original texts in different languages, for which Johansson (2007: 9) suggests the term *comparable corpus*. In such a corpus the texts in each language have been selected according to the same criteria (genre, content, publication date etc.).

The bidirectional translation corpus model of the ENPC in fact combines the translation corpus with the comparable corpus in that the original texts are comparable (matched as far as possible for genre, publication date, length of text samples, etc.). For further details of the ENPC design see Johansson, Ebeling & Oksefjell (1999/2002) and Johansson (2007: 11 ff).

As mentioned above, the ENPC was developed in close co-operation with the ESPC team, and in the early stages also with a Finnish team preparing a similarly structured English-Finnish parallel corpus; see further Aijmer et al. (1996). Later, more parallel corpora have been compiled according to the ENPC model: for example the PLECI corpus of English and French (Poitiers-Louvain Échange de Corpus Informatisés) is the result of cooperation between the universities of Louvain and Poitiers; the English/German translation corpus has been compiled at Chemnitz University, and the ACTRES parallel corpus of English and Spanish has been compiled at the University of León.[4] At the University of Oslo, the ENPC has become part of a family of corpora, under the umbrella term Oslo Multilingual Corpus. This corpus collection comprises two bidirectional translation corpora (German-Norwegian and French-Norwegian), a three-way translation corpus of English, German and Norwegian, and a translation corpus of Norwegian originals with translations into English, German and French. A parallel corpus of Russian and Norwegian is under way as a separate project.[5]

2 Contrastive analysis

Contrastive analysis is the systematic comparison of two or more languages, with the aim of describing their similarities and differences (Johansson 2007: 1). James (1980: 3) draws a distinction between contrastive and comparative investigations by pointing out that the former is typically "concerned with a *pair* of languages", and is "founded on the assumption that languages can be compared". Contrastive analysis "involves two steps: *description* and *comparison*; and the steps are taken in that order" (James 1980: 63).

Since it is obviously impossible to compare whole languages in a single investigation, a contrastive study will consist in comparing a limited number of linguistic phenomena across (at least) two languages. The items to be compared across languages are selected on the basis of *perceived similarity* (Chesterman 1998: 54), which can be formal or functional and involve all aspects of language, such as phonological features, lexical items, grammatical categories, and discourse phenomena. The investigation proceeds through various steps of forming and testing hypotheses about the degree of similarity between the items under comparison until a satisfactory description is arrived at (ibid.).

In order to ensure the validity of such an undertaking, a contrastive analysis also needs a *tertium comparationis*, i.e. a measure by which we can be fairly certain we are comparing like with like. A frequently suggested *tertium comparationis* is translation equivalence, which implies that the items in the two languages convey (more or less) the same meaning. (e.g. James 1980: 178, Chesterman 1998: 29 ff, Johansson 2007: 3).

On this background the usefulness of a bidirectional corpus for contrastive analysis is obvious: it provides an in-built *tertium comparationis* through translation equivalence and text comparability. "The paired texts reveal the interlingual identifications made by translators, and the use of parallel corpora containing such texts could be regarded as the systematic exploitation of the bilingual intuition of the translators whose work is represented in the corpora" (Johansson 1999: 117).

Other advantages of multilingual corpora for contrastive analysis, apart from the ready access to (relatively) large quantities of bilingual data have been summed up nicely by Aijmer and Altenberg (1996: 12):

- They give insights into the languages compared – insights that are likely to be unnoticed in studies of monolingual corpora.
- They can be used for a range of comparative purposes and increase our understanding of language-specific, typological and cultural differences, as well as of universal features.
- They illuminate differences between source texts and translations, and between native and non-native texts.
- They can be used for a number of practical applications, e.g. in lexicography, language teaching, and translation.

Johansson (1999: 117) furthermore points out that "while contrastive studies in the past were particularly concerned with a comparison of (parts of) language systems in the abstract, corpora now provide us with the tools for comparing languages in use."

3 A methodology for using parallel corpora in cross-linguistic studies

A bidirectional translation corpus such as the ENPC is ideally suited for investigating a word, phrase or construction in one of the languages to see what it corresponds to in the other language. Almost irrespective of the search term, the output of the search will give a range of corresponding constructions, which can be referred to as a *translation paradigm* (see further below and Johansson 2007: 23 ff). For the corresponding constructions, the term 'correspondence' was chosen in preference to such alternatives as source/translation ('correspondence' covers both) or equivalent (correspondences are not necessarily equivalents outside specific contexts).

Figure 2, taken from Johansson (2007: 25) outlines a system network for cross-linguistic correspondences, as they can be uncovered by a parallel corpus investigation. Depending on the direction of translation, correspondences are translations or sources. In most cases translations are overt, i.e. there is a linguistic expression corresponding to the search term, but zero correspondence also occurs, e.g. where a word or phrase has been omitted in the translation or the translator has added something which is not explicit in the source. Finally, an overt correspondence can be congruent or divergent. In the former case the correspondence belongs to the same grammatical class as the search term, and in the latter it does not.

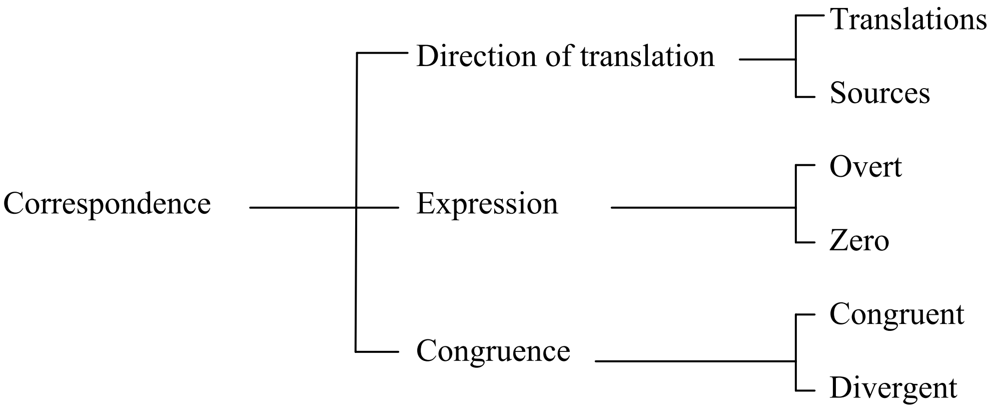

Figure 2. A system network for types of cross-linguistic correspondence (Johansson 2007: 25)

Examples (2)-(4), from the English-Swedish Parallel Corpus (ESPC), illustrate the range of correspondences. The search term was the Swedish linking adverb *emellertid*, meaning roughly 'however'. In (2) the correspondence is indeed *however*; thus it is congruent. In (3) the correspondence of *emellertid* is *but*, which clearly preserves the meaning of *emellertid,* but belongs to a different part of speech, and is thus a divergent correspondence. Finally, example (4b) contains no word that reflects the meaning of *emellertid*, thus the correspondence is zero.

(2a) Herr Cohn hade *emellertid* inga anlag för sådan förtröstan. (SCO1)

(2b) Mr Cohn, *however*, had no aptitude for that kind of consolation. (SCO1T)

(3a) Denna oktoberkväll var Sidner *emellertid* tyst, … (GT1)

(3b) *But* that October evening Sidner was quiet. (GT1T)

(4a) Militär verksamhet står *emellertid* också för en omfattande miljöförstöring. (ETHE1)

(4b) Military activity is also responsible for widespread environmental destruction. (ETHE1T)

Table 1, based on Altenberg's (1999) study of connectors in English and Swedish, shows the Swedish translation paradigm for *however,* and the English translation paradigm for the most frequent Swedish correspondence of *however*, namely *emellertid.* As shown by Table 1, members of a translation paradigm can be congruent or divergent, overt or zero.

Table 1. Translation paradigms for *however* and *emellertid* (Altenberg 1999: 259 f.)

Swedish translations of *however* (N=109)	English translations of *emellertid* (N=103)
emellertid (51 = 47%) *men* ('but') (36 = 33%) *dock* 'but'/'though' (14 = 13%) *ändå* 'still' (2) *däremot* 'on the other hand' (1) *i alla fall* 'in any case'(1) Ø (4)	*however* (83 = 81%) *but* (3) *yet* (3) *anyway* (1) Ø (13)

Paradigms of correspondence highlight the fuzzy borderlines between lexis and grammar and grammar and discourse. A good example are modal verbs, which will typically have a wide range of correspondences representing not only different parts of speech, but entirely different wordings. For instance, Løken (2007: 72) identified the following translation paradigm for the Norwegian modal *kan* ('can'):

Modal aux: *can, could, may, might, 'll, will, would, should*

Other verbs: *know, enable, have, have to, had better*

Adjectives: *possible, able, capable.*

Adverbs: *maybe, perhaps*

Suffix: *-able*

Example (5) illustrates that *kan* is rendered by a modal adverb in the translation, conveying the same modal meaning of epistemic possibility. Other types of correspondence may involve greater syntactic differences between the source and the translation, e.g. if *kan* corresponds to *it is possible* or *somebody is (cap)able*.

(5a) Valget av tidspunkt **kan** også inneholde et stenk av egoisme. (KH1) [lit: the choice of time-point can also contain a touch of egotism][6]

(5b) **Maybe** his choice of timing also contained a touch of egotism. (KH1T)

Altenberg (1999) suggests a method for measuring the degree to which correspondences are translation equivalents. The method presupposes a bidirectional parallel corpus. Mutual correspondence is calculated and expressed as a percentage by means of the formula

$$\frac{(A_t + B_t) \text{ x } 100}{A_s + B_s}$$

"A_t and B_t are the compared categories or items in the translations, and A_s and B_s the compared categories in the source texts. The value will range from 0% (no correspondence) to 100% (full correspondence)" (Altenberg 1999: 254). Using the translation paradigms from Table 1 as a starting point, we can calculate the mutual correspondence of *however* and *emellertid*: (51 + 83) x 100 / (109 + 103) = 63.2. The MC of *however* and *emmelertid* in the ESPC is thus quite high, a finding which can be reassuring e.g. for lexicographers and translators.

4 The use of bidirectional parallel corpora for research

Like monolingual corpora, multilingual corpora are particularly well suited for studies of lexis, lexico-grammar, and discourse features that can take lexis as their starting point. A tagged version of the corpus, however, makes it easier to study grammatical constructions.

A broad range of phenomena have been and are being investigated, e.g. the use of individual verbs (e.g. Ebeling's (2003) study of the Norwegian verbs *bli* and *få* and their correspondences in English, and Viberg's series of articles on Swedish verbs in comparison with other languages (e.g. 2005)), modality (e.g. Aijmer 1998, Løken 2007), particular syntactic constructions (e.g. Ebeling's (2000) study of presentatives in English and Norwegian and Johansson's study of clefts in English and Swedish), connectives (e.g. Altenberg 1999), sentence openings (e.g. Hasselgård 2004, Johansson 2007) and other discourse phenomena such as discourse markers (e.g. Aijmer and Simon-Vandenbergen 2006) and information density (Fabricius-Hansen 1998).

Although multilingual corpora clearly offer great possibilities for contrastive research, they also have their limitations. As with corpus studies in general, one can only search for something that is explicit in the text. The ENPC has been part-of-speech tagged, but not parsed (syntactically annotated), i.e. it is not possible to get directly at grammatical constructions, patterns of word order etc. However, some of these problems can be overcome by identifying typical (and searchable) expressions of a grammatical constructions, e.g. presentatives, clefting, phrasal verbs, inversion. With the tagged corpus, one can also use a combination of part-of-speech tagging, filters and wildcards to find the relevant constructions. In any case researchers who want good recall (i.e. to find *all* the relevant constructions in the corpus) will probably need to be prepared for rather a lot of work tidying up the search results. Certain phenomena, such as sentence openings, subject selection and information structure, simply cannot be retrieved automatically from a corpus that is not annotated for the relevant features. In studying such phenomena, one needs access to running text. It is still an advantage to use texts from an electronic corpus for this type of investigation in that the material will be available to the rest of the linguistic community and in the relatively easy access to supplementary material in case only a selection of corpus texts have been manually analysed.

Some special limitations apply to translation corpora. An obvious one is that they are restricted to texts and text types that have been translated. This precludes the study of many types of text, such as conversation, daily newspapers, and academic prose (the latter two text types are at least very rarely translated between English and Norwegian). Cross-linguistic investigations of non-translated text types thus have to rely on comparable corpora.[7] A translation corpus that aims to be reasonably recent, representative and balanced will probably never be very large, at least not if one of the languages concerned is not a major world language. The size of the corpus will thus restrict studies of less frequent lexical and grammatical constructions. To some extent, these limitations can be overcome by supplementing the corpus material with data from other sources, such as comparable monolingual corpora and elicitation experiments.

Finally, a problem of using translation corpora for contrastive studies is that there may be faulty or less successful translations in the material. See e.g. Mauranen (1999) for a discussion of this problem. If the translation corpus is bidirectional, one can control for translation effects by comparing original and translated text in the same language. But if the translation corpus is unidirectional, one needs to take into account that translated texts are coloured by the source language and by the translation process itself (see e.g. Steiner 2004).

5 Case study: *be going to* and *komme til å*

5.1 Introduction

As an example of how a bidirectional parallel corpus can be used in contrastive analysis, I will undertake a comparison of two future-referring expressions in English and Norwegian, namely *be going to* and *komme til å* ('come to$_{prep}$ to$_{inf.\ marker}$') to investigate the extent to which they are equivalent. Both expressions consist of a motion verb followed by an infinitive, thus showing a formal similarity. Both are described in grammars as common expressions of the future, though they are less common than the competing expressions with modals (*will* in English and *skal* ('shall') in Norwegian).

Be going to is described by Quirk et al. (1985: 214) as the 'future fulfilment of the present' or the result of present intention or present cause. Similarly, Huddleston & Pullum (2002: 210) associate it with present intention or arrangement. They also note that the past tens form *was going to* quite often has 'an implicature of non-actualisation' (2002: 211). Declerck (2006: 107) distinguishes two meanings of *be going to*: 'futurish', which is linked to a present situation, and 'future tense', simply expressing future time reference. In the latter case, "the only difference [from the *will*-future] is that *be going to* is less grammaticalized as a marker of future tense than *will* is, since it is more frequently found with predominantly present time reference" (2006: 107).

Komme til å is also linked to the present; the speaker predicts what will happen based on his knowledge at the moment of speaking, according to Faarlund et al. (1997: 543 f.). As such the construction is said to have (epistemic) modal meaning. Vannebo (1979: 259) suggests that the choice between the future tense auxiliaries *vil, skal* and *komme til å* might have to do with the nature of the speaker's knowledge, though unfortunately he does not pursue this idea.

The similarity between the two constructions is illustrated in examples (6) and (7) where they occur as each other's (congruent) translations. Both of them can be said to predict a situation fairly confidently on the basis of present knowledge, i.e. that ‘he will say something’ (6) and ‘something will happen’ (7).

(6a) I know what he’**s going to say** even before he says it. (FW1)

(6b) Jeg vet hva han **kommer til å si** selv før han sier det. (FW1T)

(7a) Ingen av dem visste hva som **kom til å skje**. (TTH1)

(7b) Neither of them knew what **was going to happen**. (TTH1T)

However, (8) has a divergent correspondence, in which the intentionality in *be going to* is rendered by the Norwegian *meningen* (‘the intention’). We may note that the example has the past tense and carries the ‘implicature of non-actualisation’ mentioned by Huddleston & Pullum (2002: 211). It appears that the Norwegian past tense form *kom til å* does not have this implicature; it would not work as a translation in (8).

(8a) "I **was going to wait** until another time we met, but I may as well tell you now. (AH1)

(8b) **Meningen** var å vente til en annen gang, men jeg kan like godt si det nå. (AH1T) [lit: ‘the intention was to wait until an other time, but I can as well say it now’]

It should be noted that *komme til å* V can be ambiguous: it can mean ‘accidentally V’, as in (9), or have an ingressive use ‘was led to V’, or ‘grew to V’ (Vannebo 1979: 264). The latter meaning is associated with the past tense and is exemplified in (10). Neither of these meanings are part of *be going to*.

(9a) Kanskje hun **kom til å svelge** dem ved et uhell? (LSC1)

(9b) Maybe she **happened to swallow** them by accident? (LSC1T)

(10a) Og siden ble det jeg som **kom til å se** mest til henne. (EHA1) [lit: 'and then was it I who came to see most to her']

(10b) And then I became the one who **ended up seeing** her most often. (EHA1T)

5.2 The corpus investigation

The entire ENPC was used for the study of *komme til å* and *be going to*, but no distinction has been made here between fiction and non-fiction. Searches were made for all forms of the expressions. The search string for English was simply *going to*, and for Norwegian all forms of *komme* followed by *til*. In both cases the material had to be tidied up manually to weed out all the cases where *to/til* was a locative preposition. No attempt was made at weeding out the occurrences of *komme til å* that do not mark future time reference, as exemplified in (9) and (10).

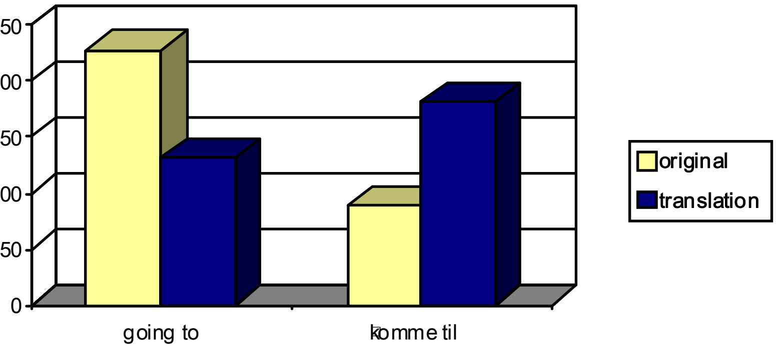

Figure 3. *be going to* and *komme til å* in ENPC fiction (raw frequencies)

Figure 3 shows the raw frequencies of *be going to* and *komme til å* in original and translated texts in the ENPC. A first observation is that *be going to* is more than twice as common as *komme til å* in original texts. There is a clear translation effect in both directions of translation in that *be going to* is less common in translations than in original texts while *komme til å* is more common in translations than in original texts. In other words, translations in both directions can be assumed to be coloured by the source texts.

To test whether the frequency differences between originals and translations is really due to the frequencies of *be going to* and *komme til å* in original texts, it is necessary to look at the correspondences of both expressions. Figures 4 and 5 show the results of this investigation.

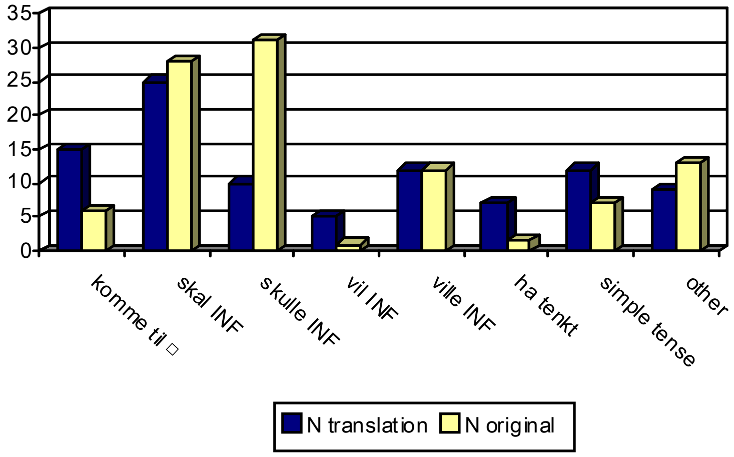

Figure 4. Correspondences of *be going to* (percentages)[8]

As shown in Figure 4, the most common sources of *be going to* are expressions with the modals *skal* and *skulle* plus infinitive (the Norwegian cognates of *shall* and *should*), followed by similar expressions with *ville* ('would'). *Skal* plus infinitive is also the most common translation of *be going to*, which suggests an overlap between different expressions of future time; i.e. it supports Declerck's (2006) interpretation of *be going to* as a neutral marker of future time reference in many of its uses. *Komme til å* is far from being the most common source of *be going to*, which indicates that the two expressions do not fit into all the same context in spite of similarities in meaning.

The correspondences of *komme til å* (Figure 5) shows a similar picture in that the most common sources and translations are expressions with *will* and *would*, thus suggesting that *komme til å* can also be interpreted as a mere marker of future time. However, *be going to* is also a frequent source.

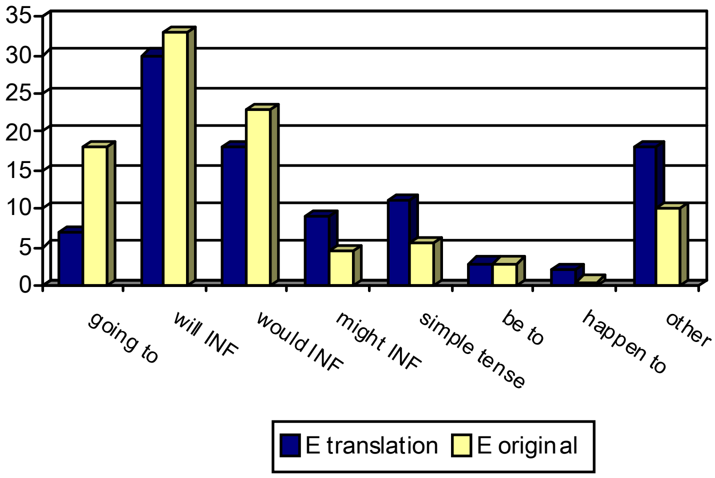

Figure 5. Correspondences of *komme til å* (percentages)

On the basis of the frequencies shown in Figures 4 and 5, we can calculate the mutual correspondence of *be going to* and *komme til å* by means of the formula given above (section 3, see Altenberg 1999: 254). Contrary to what I had expected when undertaking the investigation, the mutual correspondence of the two expressions is rather low: 12.6%. However, Figures 4 and 5 indicate that the correspondence is asymmetrical: 15% of *be going to* are translated as *komme til å* and 7% of *komme til å* are translated as *be going to*. This suggests that the meanings of the two expressions do not overlap fully; i.e. that some meanings of *komme til å* are not covered by *be going to* and vice versa. We saw above that *komme til å* can have meanings of 'accident' and 'ingression'. Further examples of these are given in (11) and (12), but it should be emphasized that these meanings are not very frequent in the material. In (11a) the verb phrase is modalized by means of *kan* ('can') denoting epistemic possibility. The modal and the future meaning are both encapsulated in the correspondence *might*.

(11a) Hun kjenner at hun er søvnig, at hun **kan komme til å sovne** mot fars jakke, hun vil ikke det. (BV2) [lit: '... that she can come to to fall-asleep against father's jacket ...']

(11b) She feels that she is sleepy, that she **might fall asleep** against father's jacket, but she doesn't want to do that. (BV2T)

(12a) … og at den kvinnen jeg leter efter egentlig var et barn den gangen hun **kom til å bety** noe for meg.“ (FC1)

(12b) … and that the woman I'm searching for was really a child when she **came to mean** something to me. (FC1T)

Example (12) has been included because it shows an English construction corresponding word for word to the Norwegian *komme til å* with ingressive meaning. The expression *come to* V occurs as a translation in (12), but was also found in English original, e.g. (13), where it happens to get a different translation.

(13a) Gradually we **came to know** each other. (ABR1)

(13b) Gradvis **ble** vi **kjent** med hverandre. (ABR1T) [lit: 'gradually got we acquainted with eachother']

5.3 Discussion of differences between komme til å and going to

The 'present cause/intention' meaning seems to work differently for the two expressions. Example (14) refers to a situation where the subject has no knowledge of the future course of action, as is evident from the question. This seems to preclude the use of *komme til å* in the translation. However, in (15) *komme til å* is used in a situation where the subject certainly does not intend to be frightened, but the speaker predicts with great certainty, presumably based on present knowledge, that she will be.

(14a) What are we **going to do**, says Ruth, … (BV2T)

(14b) Hva **skal** vi **gjøre**, sier Rut …(BV2)

(15b) Hun **komme**r bare **til å bli** redd." (THA1)

(15b) She**'ll** only **be** frightened." (THA1T)

(16a) "Are you **going to run** a hotel?" enquired Frederick reasonably, … (DL1)

(16b) "Har dere **tenkt å drive** hotell?" spurte Frederick fornuftig, … (DL1T)

Example (16) is interesting in that the use of *komme til å* would change the meaning in such a way as to raise the degree of modal certainty from median to high (see Halliday & Matthiessen, 2004: 620). That is, while the wording in (16b) is a question about plans, the corresponding question with *komme til å* would be a question about future activity. (In this context, both would probably work as translations of (16a), incidentally.) The examples thus indicate that intentionality is not part of the meaning of *komme til å*, in contrast to *going to* (cf. Quirk et al. 1985: 214). A further illustration of this is given in (17), where the future reference concerns plans, and *komme til å* would be infelicitous, as it would suggest destiny rather than intention. Example (18), which comes from a monolingual corpus of Norwegian, indicates that the future course of actions is not something the subject is planning (in which case *skal* would have been used), but simply how he envisages the future. If (18) were to be translated into English, *going to* would probably not be the best rendering of *komme til å*, since it would suggest either intention or that the speaker will be a father pretty soon, neither of which is implied by the Norwegian.

(17a) “Jeg **skal bli** tegner.” (BV1) [lit: ‘I shall become illustrator’]

(17b) “I'm **going to be** an illustrator or a painter.” (BV1T)

(18) Jeg **kommer til å bli** en helvetes far, vet du! Autoritær og jævlig. Nei, jeg tror jeg **kommer til å bli** en meget god far, akkurat passe streng og akkurat passe kjærlig. (Oslo-korpuset: AV/DN96/01) [lit: ‘I come to be a hell’s father, know you. Authoritarian and horrible. No, I think I come to be a very good father, just suitably strict and just suitably loving’]

A complicating factor in the comparison of future-referring expressions in English and Norwegian is that the Norwegian modal auxiliaries have tense, in contrast to the English ones. Thus, the past tense forms *ville* and *skulle* (which may be used in constructions competing with *komme til å*) can refer to past time. We saw in Figure 5 above that *skulle* + infinitive was a very common source of *be going to*, presumably because the *going to* construction carries tense. An example of this type of correspondence is shown in (19). Note that the Norwegian modal unambiguously denotes future in the past; the deontic meaning of obligation carried by *should* would be expressed by a different modal in Norwegian (*burde*).

(19a) Han **skulle kjempe** mot meksikanerne, og jeg tilbød ham hjelp. (SH1) [lit: ‘he should fight against the Mexicans …’]

(19b) He **was going to fight** the Mexicans and I offered him aid. (SH1T)

Figure 6 surveys the forms of *be going to* and *komme til å* in the ENPC. As I had suspected, the past tense form of *going to* is more common in translated than in original English, which is probably due to the translation of Norwegian past-tense modals; the past tense gives *be going to* a competitive edge over the *will*-future. Incidentally, the present tense *be going to* was found to occur to a great extent in direct speech, which may be due to the preponderance of fictional material using the past tense in the narrative sections. It is also noteworthy that *komme til å* is quite frequently modalized, while this does not happen at all with *going to* in the ENPC.[9] If the *komme til å* is modalized, the correspondence omits either the modal or the future-referring expression, as shown in (20) below.

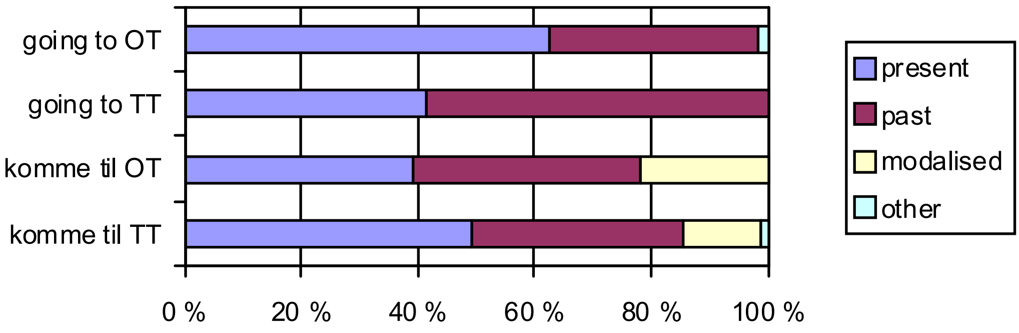

Figure 6. Forms of *be going to* and *komme til å* (OT = original text; TT = translated text)

(20a) Det som likevel synes å ha hatt størst umiddelbar virkning, var ryktene om at Sovjet **ville kunne komme til å foreslå** en slags finsk pakt med Norge. (GL1) [lit: '…were the rumours about that Soviet would could come to to suggest a type Finnish pact with Norway']

(20b) What seem to have had the most immediate effect, however, were rumors that the Soviet Union **might suggest** a type of Finnish pact with Norway. (GL1T)

It may be because of the great certainty of the prediction in *komme til å* that the expression needs to be modalized when the situation is hypothetical. (Note, incidentally, that Norwegian modals can follow each other in the same verb phrase, thus creating further problems for the translator!) Unfortunately, the further exploration of the correlation between verb forms and translation correspondence will have to await a future study.

5.4 Komme til å *vs.* going to*: summary of findings*

Having carried out this small investigation of two future-referring expressions, we can conclude that English *be going to* and Norwegian *komme til å* have some formal similarities and also share the core meaning of future-time reference. However, they also differ in a number of respects. The corpus investigation showed that they differ in frequency of occurrence, with *be going to* being over twice as frequent as *komme til å* in original texts. This may be an indication that *komme til å* is a more marked choice as a future-time referring expression than *be going to* in relation to competing expressions. Correspondence types will have to be correlated with tense forms. It seems that *be going to* is closer to a neutral future meaning than *komme til å;* i.e. it is further grammaticalized as a future tense, cf. Declerck (2006).

Table 2. Survey of meanings of *going to* and *komme til å*

	going to	*komme til å*
Prediction based on present cause/knowledge	+	+
Prediction based on present intention	+	?
Envisaged, non-intended future	?	+
Non-actualization (past tense)	+	-
Ingressive	-	+

In addition to referring to the future, both expressions carry with them additional meanings associated with intention, modal certainty and actualization. Table 2 is an attempt at summing up the differences.

The two expressions can both be used in contexts where the speaker makes a prediction grounded in the present, such as present cause, evidence or knowledge. As shown in examples (14)-(20) the expressions differ as to the implication of plan and intention. Thus, if the notion of plan/intention is important, *komme til å* does not work, but *going to* is fine. On the other hand, if a future situation is envisaged that is not part of a plan and does not have a clear grounding in the present, *going to* is not an obvious choice of future marker, but *komme til å* favours this type of context. The notion of non-actualization associated with the past tense of *be going to* does not apply to *komme til å*. The ingressive meaning of *komme til å* exemplified by (10) above is not shared by *going to*. In addition to the meanings included in Table 2, *komme til å* has an additional use which is not a future marker, namely that of 'accidentally V', as exemplified in (9).

As regards the correspondences of *be going to* and *komme til å* we have seen (Figures 4 and 5) that the most common sources and translations of both expressions are the most frequent future time referring expressions (*will/would* INF and *skal/skulle* INF). This may be due to the tendency of translators to normalize, i.e. to choose less marked forms in the translation. Conversely, some of the correspondences are more lexically explicit than *komme til å/going to*, e.g. *ha tenkt å / intend to* (subject's intention); *was to* ('was led/destined to'). These are examples of explicitation, and may be handy when an implicature of *komme til å/going to* (cf. Table 2) is crucial to the interpretation of the sentence. Finally, *be going to* may be needed in translation for syntactic reasons, as English modals lack non-finite forms and do not show tense clearly.

The picture of correspondence is a complex one, in spite of the rather similar descriptions in grammars of *be going to* and *komme til å.* Syntactic differences between *will/skal*-future expressions may go some way towards explaining the difference in distribution. However, the study of correspondences unearthed some subtle differences of meaning regarding speaker certainty and present cause/intention. As indicated by findings in Figure 6, a further study of future-time referring expressions will have to take both tense and modality into account in a more systematic fashion than has been done here, and ideally also look more closely into the meanings and distribution of competing expressions of future time.

6 Concluding remarks

It will be clear from the preceding that multilingual corpora can enhance contrastive studies in a number of ways, first and foremost by ensuring that observations are based on authentic language use. The use of quantitative data, often regarded as a hallmark of corpus methodology, provides insights into patterns of usage and preferred ways of putting things in individual languages. Frequency is indeed one of the parameters along which the languages compared may differ; for example languages as closely related as English and Norwegian often have similar lexicogrammatical resources at their disposal, but use them with different frequencies and thus different degrees of markedness, cf. the present case study of future-referring expressions.

Translation corpora have a particular advantage in that they automatically provide a *tertium comparationis* for the investigation, thereby enhancing its validity. Translation corpora can yield paradigms and patterns of correspondences, which often reveal meanings and nuances of the compared terms that might have passed unnoticed in a monolingual investigation. Moreover, the translation paradigms highlight that the same meaning may be expressed by means of different linguistic categories. If the translation corpus is bidirectional it provides a means of controlling for translation bias in that originals texts in both languages can be compared. If the corpus is sampled and representative it also provides a control for the idiosyncrasies of individual authors/translators.

In brief, multilingual corpora provide a wonderful tool for contrastive analysis. However, no matter how well designed the corpora are and how sophisticated the search tools are, they do not carry out the contrastive analysis. For this, we still need the human mind. In the words of Stig Johansson:

> The importance of multilingual corpora extends beyond contrastive studies. It is up to the user to define fruitful research questions and use the corpora creatively. In this process we learn not only about individual languages and their relationships, about translation and foreign-language acquisition, but also about language in general – provided that the study becomes truly multilingual. Seeing through corpora we can see through language. (2007: 316)

References

Aijmer, K. 1998. Epistemic predicates in contrast. In S. Johansson and S. Oksefjell (eds), 277-296.

Aijmer, K. & B. Altenberg. 1996. Introduction. In K. Aijmer, B. Altenberg, M. Johansson (eds). *Languages in Contrast.* Lund University Press, 11-16.

Aijmer, K. & A.-M. Simon Vandenbergen (eds). 2006. *Pragmatic Markers in Contrast.* Oxford: Elsevier.

Altenberg, B. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In H. Hasselgård & S. Oksefjell (eds), *Out of Corpora.* Amsterdam: Rodopi, 249-268.

Chesterman, A. 1998. *Contrastive Functional Analysis*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Declerck, R. 2006. *The Grammar of the English Verb Phrase, Vol. 1*. Berlin: Mouton de Gruyter.

Ebeling, J. 1998. The Translation Corpus Explorer: A browser for parallel texts. In S. Johansson & S. Oksefjell (eds), 101-112.

Ebeling, J. 2000. Presentative constructions in English and Norwegian: a corpus-based contrastive study. PhD dissertation, Faculty of Humanities, University of Oslo, UniPub.

Ebeling, S. O. 2003. The Norwegian verbs *bli* and *få* and their correspondences in English: a corpus-based contrastive study. PhD dissertation, Faculty of Humanities, University of Oslo, UniPub.

Faarlund, J. T., S. Lie, K. I. Vannebo. 1997. *Norsk Referansegrammatikk*. Oslo: Universitetsforlaget.

Fabricius-Hansen, C. 1998. Informational density and translation. In S. Johansson & S. Oksefjell (eds), 197-234.

Filipović, R. 1969. *The Yugoslav Serbo-Croatian-English Contrastive Project*. Zagreb University, Institute of Linguistics. (Available at www.eric.ed.gov/PDFS/ED121083.pdf, accessed 30.10.2010.)

Fløttum, K., T. Dahl & T. Kinn. 2006. *Academic Voices – Across Languages and Disciplines.* Amsterdam: John Benjamins.

Halliday, M.A.K. & C.M.I.M. Matthiessen. 2004. *An Introduction to Functional Grammar. Third Edition.* London: Arnold.

Hasselgård, H. 2004. Thematic choice in English and Norwegian. *Functions of Language* 11:2. 187-212.

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

Huddleston, R. and G. K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

James, C.. 1980. *Contrastive Analysis.* London: Longman.

Johansson, M. 2002. Clefts in English and Swedish: A contrastive study of IT-clefts and WH-clefts in original texts and translations. PhD dissertation, Lund University.

Johansson, S. 1999. Corpora and contrastive studies. In P. Pietilä & O-P. Salo (eds), *Multiple Languages – Multiple Perspectives.* AFinLA Yearbook 1999 / No. 57, 116-125.

Johansson, S. 2007. *Seeing through Multilingual Corpora.* Amsterdam: Benjamins.

Johansson, S., J. Ebeling and S. Oksefjell. 1999/2002. English-Norwegian Parallel Corpus: Manual. www.hf.uio.no/ilos/english/services/omc/enpc/ENPCmanual.pdf (accessed 30.10.2010)

Johansson, S. & S. Oksefjell (eds). 1998. *Corpora and Cross-linguistic Research. Theory, Method, and Case Studies*. Amsterdam: Rodopi.

Johansson, S. & Hofland, K. 1994. Towards an English-Norwegian parallel corpus. In U. Fries, G. Tottie and P. Schneider (eds), *Creating and Using English Language Corpora*. Amsterdam: Rodopi, 25-37.

Leech, G.N. 2010. Corpus linguistics. In K. Malmkjaer (ed.), *The Routledge Linguistics Encyclopedia, Third Edition*. London: Routledge, 103-113.

Løken, B.H. 2007. *Beyond Modals: A Corpus-based study of English and Norwegian Expressions of Possibility.* PhD dissertation, Faculty of Humanities, University of Oslo, UniPub.

Mauranen, A. 1999. Will 'translationese' ruin a contrastive study? *Languages in Contrast* 2:2, 161–185.

Nordrum, L. 2007. English Lexical Nominalizations in a Norwegian-Swedish Contrastive Perspective. PhD dissertation, Göteborg University.

Quirk, R., S. Greenbaum, G. Leech, J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Steiner, E. 2004. *Translated Texts: Properties, Variants, Evaluations*. Frankfurt am Main: Lang.

Svartvik, J. 1992. Corpus linguistics comes of age. In J. Svartvik (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin: Mouton de Gruyter, 7-13.

Vannebo, K. I. 1979. *Tempus og tidsreferanse*. Oslo: Novus.

Viberg, Å. 2005. The lexical typological profile of Swedish mental verbs. *Languages in Contrast*, 5:1, 121–157.

Corpora used

British National Corpus: www.natcorp.ox.ac.uk/

English-Norwegian Parallel Corpus: www.hf.uio.no/ilos/english/services/omc/enpc/

English-Swedish Parallel Corpus: www.sol.lu.se/engelska/corpus/corpus/espc.html

Oslo Corpus of Tagged Norwegian Texts: www.tekstlab.uio.no/norsk/bokmaal/english.html

Notes

[1] I had the honour of giving a plenary lecture at the UCCTS2010 conference as a replacement for my former teacher and colleague Stig Johansson, who sadly passed a few months before the conference. He was a pioneer in corpus linguistics, starting at Lancaster University in the 1970s, where he helped complete the LOB corpus. Later, he also became a pioneer in

multilingual corpus development, being the initiator and project leader of the English-Norwegian Parallel Corpus. This paper is dedicated to his memory.

[2] The first attempt at a translation corpus was in fact made in former Yugoslavia as early as the 1960s, with parts of the Brown corpus being translated into Serbo-Croatian and plans for the compilation of a corresponding Serbo-Croatian corpus to be translated into English (Filipović 1969).

[3] A more fine-grained text type distribution would have been desirable. However, the limitations as to the types and number of texts that are translated between English and Norwegian made it impossible to collect enough non-fiction to subdivide the category. Non-fictional text types represented in the corpus are diverse and include for example biography, popular science, legal texts and tourist information. See further www.hf.uio.no/ilos/english/services/omc/enpc/.

[4] More information about these corpora can be found at the following websites: PLECI www.uclouvain.be/en-cecl-pleci.html; English/German www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/transcorpus/; ACTRES http://actres.unileon.es/.

[5] For the Oslo Multilingual Corpus, see further www.hf.uio.no/ilos/english/services/omc/, and for the Russian-Norwegian-English project, see www.hf.uio.no/ilos/english/research/projects/run/corpus/.

[6] Where the wording in the Norwegian version of the examples differs from that of the English, a literal translation is provided. Examples that are not followed by a literal rendering can safely be assumed to follow the pattern of the English corresponding sentence very closely.

[7] A good example of this is the KIAP project (Cultural identity in academic prose) at the University of Bergen; see http://kiap.uib.no and Fløttum et al. (2006).

[8] The Norwegian correspondences of *be going to* other than *komme til å* can be literally translated as follows: *skal* - 'shall', *skulle* - 'should', *vil* - 'will', *ville* - 'would', *ha tenkt å* – 'have thought to' ('intend'). The category of 'other' are those correspondences that occurred less than five times in the material.

[9] Some examples of modal + *be going to* were found in the British National Corpus, especially *might be going to*, but even they were rather rare (20 hits in 100 million words).