Paper submitted to The XXVIII International Conference on Cross-Language Studies and Contrastive Linguistics. 15 - 17 December 1994, Rydzyna, Poland

The English-Norwegian Parallel Corpus: Introduction and Applications

Prof. Stig Johansson, Department of British and American Studies, University of Oslo
Cand.philol Jarle Ebeling, Department of British and American Studies, University of Oslo

1 Introduction

1.1 Corpora in linguistic research

The place of corpora has been controversial in linguistics. Yet there is now more interest than ever in the use of corpora in linguistic research. This is connected with the increasing interest among linguists in studying language in use rather than language systems in the abstract. It is no doubt also due to the development of corpora in machine-readable form and of techniques for exploring and analysing such corpora.

A computer corpus provides data for descriptive studies. It provides a test bed for more theory-oriented work. Once compiled, it is transportable, and can be used by a large number of researchers for a range of purposes (many of which may not have been foreseen at the stage of compilation). See further the arguments in favour of corpora in Svartvik (1992: 8-10).

There is now a very large number of English computer corpora (see e.g. Taylor *et al.* 1991), and they are increasing rapidly in size as well as number. The Bank of English at Birmingham is reported to have reached (at least) 200 million words. The British National Corpus, which is now being completed, will contain 100 million words. The access to material on this scale makes it increasingly important to improve the means of exploring the corpora and to carefully consider how they can be best exploited.

One area where computer corpus work is only just beginning is the field of contrastive linguistics. There appears to be an increasing interest in the compilation and use of parallel corpora in two or more languages, and in the development of techniques for analysing such corpora (including methods of automatic alignment of parallel texts; cf. 1.4 below). The Canadian English-French Hansard Corpus is perhaps the best-known bilingual corpus. Danish researchers have compiled a trilingual corpus of legal texts (see Faber and Lauridsen 1991) and another specialised trilingual corpus containing texts in the field of biotechnology. Yet the development and use of parallel corpora for contrastive studies is in its infancy. This is the background for our work on the English-Norwegian Parallel Corpus (ENPC).

1.2 Content and structure of the corpus

The ENPC project started at the beginning of 1994, after a pilot study had been carried out in 1993, and it is expected to end in early 1996. The parallel corpus is meant to be a research tool for linguists and students interested in contrastive linguistics. It will consist of two sub-corpora: a *translation* corpus and a *comparative* corpus. The translation corpus will contain English and Norwegian originals and their translations (including both English-to-Norwegian and Norwegian-to-English translations), while the comparative corpus will consist of original texts which are matched with respect to the type of text.

Figure 1 shows the structure of the corpus and indicates the types of comparative studies which we hope to undertake (we shall come back to this point in Section 2 below). The boxes marked by broken lines represent the comparative corpus, which will be larger in size and broader as regards content than the translation corpus.

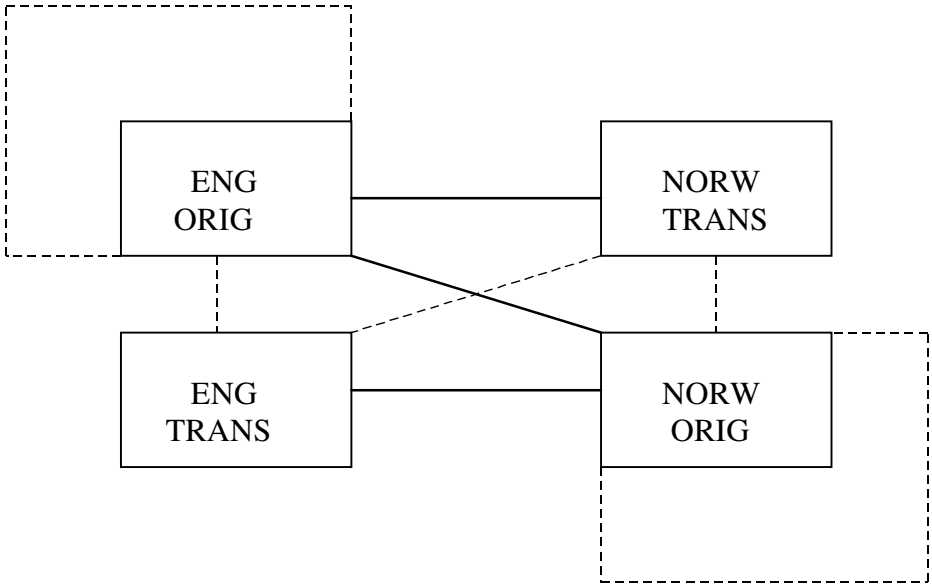

Figure 1: The structure of the Norwegian-English Parallel Corpus

The corpus will only contain written material, but both fiction and non-fictional texts will be included. To make it possible to include as many different writers and translators as possible, text extracts of 10,000 - 15,000 words will be selected and not complete books. Each text extract will start at the beginning of the book[1] and, if possible, end at a chapter boundary. This will ensure coherent if not complete texts. Where texts are very short, they will, however, be included in their entirety.

We have so far concentrated on the translation corpus and have, in close cooperation with the Department of English, University of Lund, which has a similar project going, set up a simple taxonomy to guide our text selection. See Figure 2.

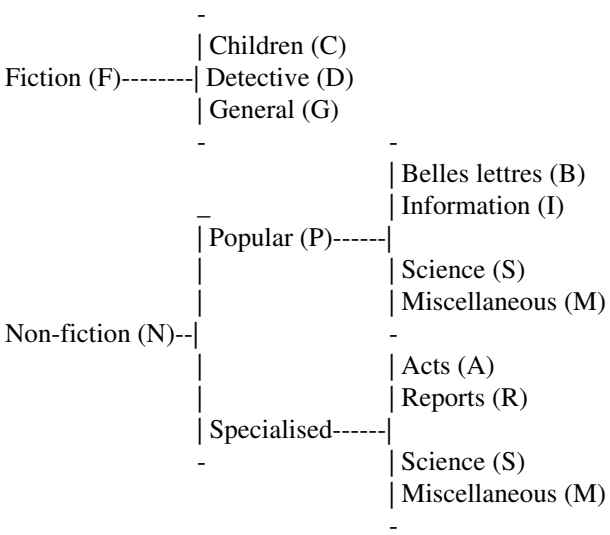

Figure 2: A simple taxonomy of text categories

[1] Front matter (bibliographic information, content, foreword, etc.) is skipped.

We hope to be able to compile as many as 160 text extracts during the first year (40 English and 40 Norwegian texts, with their translations), which will give us a corpus of approximately 1,9 million words.

Our text selection is limited by the types of texts that are translated. While a large number and a wide variety of texts are translated from English into Norwegian each year, the amount and variety of texts translated from Norwegian into English are far more limited. The text selection is also guided to some extent by ease of access, that is, we have picked books which were readily available at the university library. This means that the material so far is biased towards "serious" literature, while children's books and "kiosk" literature are under-represented. We have only incorporated fairly recent publications, from 1980 and onwards. However, for Norwegian fiction we have had to go back to 1970 to be able to get enough text extracts.

1.3 Markup

All text extracts are prepared for computer processing and are encoded according to the recommendations of the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard 1994). Markup is inserted into the text as tags. Angle brackets are used to separate tags from text: < > for 'start-tags' and </> for 'end-tags'. Tags may have attributes, e.g. for 'id' and 'link' (see example 2 below). The occurrence of tags and attributes is defined formally by a so-called 'document type definition' (DTD).

At present we are only inserting some basic markup; no grammatical parsing is done. The preparation of the electronic version of the text can be seen to comprise five steps:

1 Scan text extract and store in ASCII format;
2 Insert tags to mark paragraph boundaries ( ), page breaks (<pb n=x>) and highlighted text (<hi rend=italic/bold></hi>) such as italic or boldface;
3 Split contractions (e.g. *he'll* --> *he 'll*) and insert sentence markup (<s></s>);
4 Proofread text;
5 Check markup with an SGML-parser.[2]

All tags are inserted automatically into the text, either by simple search and replace facilities in a word processor, or by computer programs.

Example 1 shows what Gunnar Staalesen's *At Night All Wolves Are Grey* looks like before and after markup:[3]

(1)

'Because — because I *know* we're never going to get to the bottom of that case. And there's nothing that irritates a policeman more than to feel that a case has gone unsolved.'

16

'But...'

'And because,' he went on, ignoring me, 'because it was so easy to see the *pattern* in it all.'

<s>"Because ‐ because I <hi rend=italic>know</hi> we 're never going to get to the bottom of that case.</s> <s>And there 's nothing that irritates a policeman more than to feel that a case has gone unsolved."</s>
<pb n=17> <s>"But..."</s>

[2] An SGML-parser is a computer program which ensures that the markup is in accordance with a predefined description of the structure of the markup. SGML (Standard Generalized Markup Language) is an ISO standard (ISO 8879), and is the backbone of the TEI guidelines.

[3] In order to ease the alignment process, quotation marks are normalised to double quotes. Entity references (marked by & ;) are used for special characters (e.g. ‐).

<p><s>"And because," he went on, ignoring me, "because it was so easy to see the <hi rend=italic>pattern</hi> in it all."</s>

In addition to marking up the body of the text, we provide each text extract with a header (<header></header>) which contains bibliographic information about the text as well as information about dialect (BrE, AmE, etc.), text category (FD = Fiction Detective, NPB=Non-fiction Popular Belles lettres, etc.), and any special features about the text which should be pointed out, e.g. illustrations that are left out or the use of nonstandard language in the text.

The overall structure of a text extract looks like this:

<partext lang=en type=novel orig=yes id=GS1T> *Summary of header info. plus unique id*
<header> *Bibliographic info., etc.*
</header>
<text>
<front> *Left empty*
</front>
<body> *The scanned text*
</body>
<back> *Left empty*
</back>
</text>
</partext>

The amount of markup may seem overwhelming, but it is a minimum considering the kind of analyses we want to carry out (see 1.4 and 2 below). By using an international standard as basis for the markup we have the advantage of being able to exchange texts with other projects more easily; and we get access to tools which we otherwise would have to develop ourselves.

1.4 Alignment[4]

In addition to the markup shown in example 1, the text extracts in the translation corpus are aligned at sentence level (i.e. orthographic sentence). This means that a sentence in the original contains a reference to the corresponding sentence(s) in the translation. This is done by supplying each <s>-tag with a unique identifier and a link to identifier(s) of the corresponding sentence(s) in the parallel text:

(2) (DL2/T.1.s18 = Doris Lessing text 2 original/translation, chapter 1, sentence 18)

<s **id=DL2.1.s18 link='DL2T.1.s18 DL2T.1.s19'**>At once, feeling her advantage, she said, "Do n't forget you 've been living soft for four years."</s>

<s **id=DL2T.1.s18 link=DL2.1.s18**>Hun hadde fått et lite overtak og fulgte det opp.</s>
<s **id=DL2T.1.s19 link=DL2.1.s18**>"Ikke glem at du har levd godt i fire år nå."</s>

Alignment is done automatically, using an algorithm based on an anchor word list (i.e. a simple bilingual lexicon), sentence length in characters, and the occurrence of proper names, and special features of punctuation (! ? — (dash)) and markup, e.g. of highlighted text. The success rate of the automatic alignment program is very high (approx 98 per cent).

[4] A more detailed description of markup and alignment can be found in Johansson og Hofland (1994).

2 Applications

It is our hope that the corpus will provide ample material for text-based contrastive studies as well as for more specialised translation studies. The solid horizontal lines between the boxes in Figure 1 above are meant to indicate comparative studies based on original and translated texts. The solid diagonal line indicates comparison of parallel original texts. These represent two complementary ways of carrying out text-based contrastive studies.

Original and translated texts are of special importance in that they show how the same meaning can be expressed in two languages. Their use for contrastive studies is limited, however, by the fact that they may reveal as much about the translators and the process of translation as about relationships between the languages involved. This is why we set up our corpus in such a way that contrastive findings based on the translation corpus can be checked against the comparative corpus of parallel original texts.

As our corpus will contain different text categories (see 1.2 above), we hope to be able to study the relationship between text categories across languages. To what extent are differences between text categories, e.g. fiction vs. non-fiction, parallel across languages?

A corpus structured like ENPC will permit a range of translation studies. How is a particular form or construction translated? Are there differences depending upon the type of text? Do correspondences differ depending upon the direction of translation (i.e. in translating from English into Norwegian and vice versa)?

A type of study which is of particular interest is the comparison of original and translated texts in the same language (indicated by vertical broken lines in Figure 1). Gellerstam (1986) studied original texts by Swedish authors and comparable texts translated from English. The comparison revealed a great deal of evidence of "translationese", i.e. deviations in the translated texts as compared with the norms of comparable original texts in the same language. We hope to carry out similar studies and will be able to go beyond Gellerstam's work in that we will have available both original texts and translations in both languages (whereas Gellerstam's study was limited to texts in one language).

The study of translationese is significant in that we cannot expect languages to remain constant. We know that there is a great deal of English influence on the Scandinavian languages, particularly in the realm of vocabulary. Through the large number of translations, English may similarly, in a less direct manner, influence the Scandinavian languages.

Yet a further line of inquiry is shown by the diagonal broken line in Figure 1. To what extent are there features in common among translated texts? For some suggestions in this area, see Baker (1993).

3 Preliminary analyses

At the moment of writing, only half a year after its beginning, our project has not progressed far enough for us to carry out any major studies; we can only present some very preliminary findings at this point. These should, however, be of some interest in showing the sorts of analyses which can be carried out using a parallel corpus.

The studies below are based on six texts translated from Norwegian into English, and twelve texts translated from English into Norwegian, in all eighteen pairs of texts, amounting to some 250,000 words for each of the languages. The texts, which are all novels, are listed in Appendix 1.

3.1 Some basic statistics

Tables 1 and 2 summarise the relationship between the texts in terms of the number of paragraphs, S-units (orthographic sentences), and words.

With a couple of exceptions, we find a very close correspondence as regards the number of paragraphs, and there is indeed little reason to suspect that there should be any marked differences in paragraphing conventions between English and Norwegian. We note, however, an overall tendency for the Norwegian texts (originals as well as translations) to contain more paragraphs. A paragraph has been defined orthographically as a passage of text marked off by indentation and/or the use of blank lines. We have not yet studied the reasons for the differences in the number of paragraphs, but suspect that it has to do with the representation of dialogue, where there is probably a good deal of individual variation. For our project, the statistics for paragraphs are mainly important in connection with our work on automatic alignment. Originally, our automatic alignment was based on a matching of paragraphs. As there are differences in paragraphing, albeit not very frequently, we gave this up, and the matching is now based on a fixed window moving through the text (defined in terms of sentences rather than paragraphs).

If we might at the outset expect the number of paragraphs to be more or less constant, there are good reasons to assume that a translation might differ from the original as regards the number of words. The figures in Table 1 show an average increase of about 1450 words (just over 10 per cent) in the English translation as compared with the Norwegian original. There are a number of reasons why we might expect there to be fewer words in the Norwegian texts, chiefly the more frequent use of compounding and the marking of definiteness of nouns by a suffix rather than by an independent word (compare: husbygging vs. house construction, the building of houses; huset vs. the house). This might be counterbalanced to some extent by the more frequent use of finite clauses in Norwegian, particularly corresponding to the more compact English ing-clauses (compare: running across the field vs. mens han løp over jordet). Notice, by the way, that English contractions like I've and can't have been split up into two words (to improve the efficiency of the automatic alignment technique). This, however, does not contribute to a larger number of words in English, as the corresponding words are not contracted in Norwegian.

Table 1: The number of paragraphs, orthographic sentences (S-units), and running words in the Norwegian originals/English translations

	Original			Translation		
	Paragraphs	S-units	Words	Paragraphs	S-units	Words
FC	191	726	14,185	148	706	14,694
GS	342	1,108	12,762	349	1,088	14,961
HW	535	1,172	13,911	468	1,120	15,796
KF	101	721	12,567	101	741	13,571
KFL	54	637	12,696	51	652	14,711
SH	405	1,260	15,165	399	1,225	16,265
Total	1,628	5,624	81,286	1,516	5,532	89,998
Mean	271	937	13,548	253	922	15,000

Table 2: The number of paragraphs, orthographic sentences (S-units), and running words in the English originals/Norwegian translations

	Original			Translation		
	Paragraphs	S-units	Words	Paragraphs	S-units	Words
AB	55	598	13,479	53	619	13,902
AW	225	986	14,219	223	987	14,507
DF	421	1,042	14,855	422	1,125	14,829
DL	432	1,234	14,874	432	1,323	17,336
FW	197	942	13,159	197	936	13,400
JA	279	720	13,601	281	749	13,452
JH	367	585	13,277	367	596	13,528
LD	408	1,195	14,380	525	1,058	13,256
MA	155	722	11,935	156	727	11,321
RD	384	966	12,471	398	1,044	13,150
RR	304	1,041	14,449	299	1,110	14,383
WG	392	1,144	13,249	391	1,174	14,022
Total	3,619	11,175	163,948	3,744	11,448	167,086
Mean	302	931	13,662	312	954	13,924

If we turn to the Norwegian translations of the English texts, we might expect a decrease in the number of words. What we do find, however, is that there is instead an average increase of about 2% (though there are differences between the individual texts, some showing an increase and others a decrease as compared with their Norwegian counterpart). We have not had time to study these matters in detail, and our material is

nevertheless too small to allow any hard and fast conclusions, but we may tentatively interpret the statistics above as a reflection of two tendencies: (1) the tendency for Norwegian texts to contain fewer words (for the reasons given above), and, possibly, (2) a general tendency for translated texts to be longer than original texts.

A comparison of the number of orthographic sentences (S-units) gives fairly consistent results, regardless of the direction of translation. With a couple of exceptions, the Norwegian texts (both original texts and translations) contain more sentences. Differences are not large, but nevertheless worth noting: the Norwegian original texts on average contain 15 sentences more than the English translations (representing a decrease of 1.5 per cent in the translation); the Norwegian translations on average contain 23 sentences more than the English original texts (an increase of 2.5 per cent). The difference in sentence division will be examined more closely in the next section.

3.2 Sentence division

There are good reasons why we might expect a difference in sentence division. English can easily attach a non-finite clause at the end of a sentence. In such cases Norwegian will either have to resort to a finite clause (coordinate or subordinate) in the same sentence or begin a new sentence. Inspecting our material, we find a good number of instances of the latter solution. These are some examples from Doris Lessing's The Good Terrorist and its Norwegian translation:

(3)
- Without referring to him she ran up the stairs and he followed slowly, listening to how she banged on doors, and then, hearing nothing, flung them open.
- Hun løp opp trappene uten å ofre ham et blikk og han gikk langsomt etter. Han hørte henne dundre på dør etter dør, og rive dem opp ettersom ingen svarte.

(4)
- Jasper drank down his cup at once, and sat looking at the thermos, wanting more.
- Jasper tømte koppen sin med det samme og satt der og så på termosen. Han ville gjerne hatt mer.

(5)
- "And if they wo n't?" he said, knowingly, sharing expertise with her, a friendly moment.
- "Og hvis de ikke vil, hva så?" sa Jasper. De delte viten og erfaringer, det var et lite, fortrolig øyeblikk.

We notice that the ing-clauses in these examples have been translated as independent sentences. Here are some other examples of differences in sentence division from the same book:

(6)
- She said, "This is Faye, I am Roberta."
- "Dette er Faye," sa hun. "Jeg heter Roberta."

(7)
- And suddenly they all knew why it was Faye they had been eyeing so nervously, Faye who had dominated everything.
- Og plutselig visste alle sammen hvorfor det var Faye de hadde holdt så vaktsomt øye med. Faye som hadde dominert hele samtalen.

(8)
- She sounded breathless: Bert took it as fear, and he said, with loud, cold contempt, "Shit-scared."
- Hun hørtes rent andpusten ut. Bert tok det for redsel, og sa med kald forakt: "De ble vettaskremte."

These examples show that a difference in sentence division may be a purely orthographic phenomenon. It appears that the English text sometimes has a weaker punctuation mark (comma, colon, semi-colon), where the Norwegian translator has inserted a full stop. In our sentence division algorithm we do not recognise a sentence boundary in the former case, only where a mark of end punctuation (full stop, exclamation mark, exclamation mark) is followed by a word with an initial capital or is at the end of a paragraph.

In examining one of our Norwegian texts, Finn Carling's Under aftenhimmelen, and its English translation we find a good number of examples of differences in the representation of dialogue, e.g.:

(9)
<s>Så fortsatte han:</s>
<s>Hvilken rørende naivitet!</s>
<s>Then he continued: What touching naïvety!</s>

In other words, the dialogue is represented in the Norwegian text by paragraphs (following our orthographic definition; cf. 3.1), whereas the English text shows no paragraph division. We have not studied the matter exhaustively, but expect that differences of this kind partially account for the findings presented above (3.1) on the number of paragraphs in the English vs. Norwegian texts.

3.3 Word correspondences

As an example of the study of word correspondences, consider the frequency of the three main coordinating conjunctions and, but, and or vs. their Norwegian equivalents og, men, and eller. See Table 3.

Table 3: The frequency of some coordinate conjunctions in the English and Norwegian texts. Relative frequencies per 100,000 words are given within parentheses.

Norwegian original		English translation		English original		Norwegian translation	
og	3,380 (4,158)	and	3,118 (3,464)	and	4,590 (2,800)	og	5,531 (3,310)
men	647 (796)	but	638 (709)	but	926 (565)	men	1,080 (646)
eller	287 (353)	or	324 (360)	or	560 (342)	eller	577 (345)

While the relative frequencies of or and eller are very similar in the four groups of texts, there are wide differences with the other two pairs. Differences are particularly striking with and vs. og. This is not unexpected. It is partially connected with the frequency of non-finite clauses in English (cf. 3.2). As they are frequently rendered by Norwegian coordinate clauses, we expect a higher frequency for the Norwegian conjunction og. Other factors contributing to a higher frequency of Norwegian og are its use in the sense of 'also' and its use in pseudo-coordinate structures like satt og leste (=sat reading, was reading), sto og stirret (=stood looking, was looking), etc.

The patterning of og vs. and is not the same in the four groups of texts. The highest relative frequency is found in the Norwegian original texts, the lowest in the English original texts. This is incidentally also the patterning we find with men vs. but; see Table 3. It is tempting to suggest that the translations produce choices which are closer to the original texts, i.e. translationese in the sense of deviations in the direction of the norms for the source language.

The correspondences can be studied in detail using the aligned corpus, and eventually the results can be checked against studies of coordination patterns in the larger comparative corpus. As a proper investigation of coordination requires an in-depth study, it will not be explored further in this context. Instead, as an additional example of a study of word correspondences, consider the use of the English verb come and its Norwegian cognate komme. See Table 4.

Table 4: The frequency of forms of the English verb come and the Norwegian verb komme. Relative frequencies per 100,000 words are given within parentheses.

Norwegian original		English translation		English original		Norwegian translation	
komme	47 (58)	come	79 (88)	come	164 (100)	komme	128 (77)
kommer	39 (48)	comes	16 (18)	comes	19 (12)	kommer	122 (73)
kommet	23 (28)	coming	13 (14)	coming	46 (28)	kommet	82 (49)
kom/kjem	226 (278)	came	136 (151)	came	178 (109)	kom/kjem	360 (215)
Total	335 (412)		244 (271)		407 (248)		692 (414)

Table 4 shows that the relative frequency is far higher for the Norwegian verb. Note that the forms cannot be compared one by one as they are listed in the table. For example, English come may correspond to all the Norwegian forms: komme (cf. to come), kommer (cf. they come), kommet (cf. they have come), and kom (cf. didn't come).

To study the correspondences more closely, we selected the Norwegian past tense form kom. Sentences containing this form were retrieved together with their English equivalents, as identified by the automatic alignment program. Because of the current state of the material, it was not possible to retrieve every occurrence. The list below identifies English verbs corresponding to kom for approximately 180 instances in the English translation material and a corresponding number in the English original texts. Forms found in both groups of texts are italicised. The verbs are given in their base form. Excluded are a number of examples, where the translation was very free and it was not possible to identify a verb form corresponding to kom.

Translations of Norwegian kom:

appear (1), arrive (9), be (6), be born (2), be going to (1), be going to get (1), begin (1), come (101), dribble out (1), drift (2), enter (3), escape (2), extend (1), follow (1), get (14), go (1), happen (1), leave (1), make one's way (1), pass (1), (be) put (2), return (3), ride (2), rise (1), run (1), say (1), sound (1), spread (1), start (3), step up (1), walk (1), will/would (3)

Original forms corresponding to Norwegian kom:

add (1), arise (1), arrive (6), be (10), be going to (2), be sure to (1), bring (3), call (1), come (76), confront (1), demand (1), drop in (1), emerge (1), enter (3), express (1), get (4), give

(2), go (5), happen (1), let out (1), reach (5), reappear (1), remember (1), return (7), ride (1), run (1), slouch (1), start (1), visit (3)

Not surprisingly, the main equivalent of kom turns out to be a form of come, but many other English verbs are represented, chiefly movement and transitional-event verbs. The range of equivalents is much higher than indicated in bilingual Norwegian-English dictionaries.

Without a detailed semantic and collocational study, preferably based on a much larger material, these results are of limited value. At this stage we can merely suggest how material of this kind can be exploited:

1. We envisage that a bilingual concordance derived from an aligned parallel corpus will be a very useful supplement to a bilingual dictionary. Here learners can study translation equivalents in context, whereas bilingual dictionaries are usually highly restricted as regards contextual specifications.

2. Bilingual concordance material for individual word pairs or for groups of semantically related words will be very useful for contrastive lexical studies.

3. Using material of the kind we have presented above for come and komme, it may be possible to define transfer probabilities in going from one language to another. Such probabilities may be of use in machine translation systems. For our own purposes, we can use such material to revise our anchor word list and, ultimately, to improve the success rate of our program for automatic alignment.

3.4 Tag questions and negation

A study of the relative frequencies in our material of English not vs. Norwegian ikke reveals a consistent difference; see Table 5. The English negator is less frequent than its Norwegian counterpart in both groups of texts. One possible reason is that negation may be expressed more frequently in English by negative pronouns and determiners (nobody, no, etc.) than in Norwegian. There are differences between English and Norwegian in the use of the two types of negation (cf. Johansson and Lysvåg 1986: 238f.), but they have not been adequately described. The corpus provides an ideal source for such a study.

Table 5: The frequency of the English negator not and the Norwegian negator ikke. Relative frequencies per 100,000 words are given within parentheses.

Norwegian original		English translation		English original		Norwegian translation	
ikke/ikkje	866 (1,065)	not/n't	800 (889)	not/n't	1,790 (1,092)	ikke/ikkje	2,210 (1,323)

One English-Norwegian difference which might produce a lower negative frequency in Norwegian, i.e. going against the tendency above, relates to tag questions. Though there is a formally somewhat related structure in Norwegian (ikke sant, literally 'not true?'), more adequate Norwegian equivalents are probably pragmatic particles like jo and vel.

We retrieved all tag questions from our English material, both original texts and translations. As the search pattern simply specified a pronoun followed by a question mark (he? it? etc.), we may have missed some occurrences, but we have no reason to suppose that the results would be biased one way or the other because of this. Needless to say, we

eliminated all examples produced by the search which were not tag questions. The results were (relative frequencies per 100,000 words are given within parentheses):

English original 47 (28) English translation 9 (10)

The relative frequency is thus far lower in the English translation. It is tempting to suggest that this is due to the fact that Norwegian lacks a syntactic equivalent to English tag questions. In other words, the lower frequency in the English translated texts may be a reflection of translationese. This again needs to be studied further, preferably within the broader context of an investigation of tag questions, pragmatic particles, and other related structures.

4 Prospects

At the current stage, we are chiefly focusing on the building of the corpus, and we have not been able to carry out any large-scale investigations. The examples given in this paper should, however, be sufficient to show the possibilities of using our corpus. We believe that contrastive studies will benefit greatly from being related to parallel corpora.

Finally, let us look briefly at future developments in our project.

4.1 Multi-dimensional expansion of the corpus

There is no reason why the corpus should stop growing at the end of our two-year project, and there is no limit to the number of texts which can be incorporated. This means that, as the two languages develop, it will be possible to log this development by adding more texts.

Our work will also be extended by cooperation with other contrastive projects. There are plans in Norway to set up a larger project where other languages, German and French, will join in. As already mentioned, we are also cooperating with a contrastive team at the University of Lund in Sweden, and we have agreed to use the same English original texts for the translation corpus. We are further setting up cooperation with contrastive projects in Denmark and Finland. In this manner we hope to produce material which is suited for multilingual comparison.

4.2 Analysis tools

The development of good automatic parsers in the last few years has convinced us that it will be possible to tag the English part of the corpus grammatically with a high degree of accuracy. For the Norwegian part of the corpus, prospects are less promising, and it will therefore take a good many years before we can produce a bilingual corpus where texts are tagged in both languages.

We will continue to improve our program for automatic sentence analysis.[5] The program has already been tested for other languages, with good results. We have plans to develop a program for automatic word alignment, as well as for sentence alignment.

Last but not least, we hope to develop software tools that are suitable for exploring and analysing parallel corpora.

[5] The computer programs for our project have been developed by Knut Hofland, Norwegian Computing Centre for the Humanities, Bergen.

References

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and technology. In honour of John Sinclair*, ed. by Mona Baker, Gill Francis and Elena Tonini-Bonelli. 233-250. Philadelphia: Benjamins

Faber, Dorrit & Karen M. Lauridsen. 1991. The compilation of a Danish-English-French corpus in contract law. In *English computer corpora: Selected papers and research guide*, ed. by Stig Johansson and Anna-Brita Stenström. 235-243. Berlin: Mouton de Gruyter

Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. In *Translation studies in Scandinavia*, ed. by Lars Wollin and Hans Lindquist. 88-96. Lund: CWK Gleerup

Johansson, Stig and Knut Hofland. 1994. Towards an English-Norwegian parallel corpus. In *Creating and using English language corpora*, ed. by Udo Fries, Gunnel Tottie and Peter Schneider. 25-37. Amsterdam: Rodopi

Johansson, Stig and Per Lysvåg. 1986. Understanding English grammar. Part 1. Oslo: Universitetsforlaget.

Sperberg-McQeen, Michael C.M. & Lou Burnard (eds.). 1994. *Guidelines for electronic text encoding and interchange*. TEI P3. Chicago & Oxford: Association for Computers and the Humanities/Association for Computational Linguistics/ Association for Literary and Linguistic Computing

Svartvik, Jan (ed.). 1992. *Directions in corpus linguistics*. Trends in Linguistics. Studies and Monographs 65. Berlin: Mouton de Gruyter.

Taylor, Lita, Geoffrey Leech & Steven Fligelstone. 1991. A survey of English machine-readable corpora. In *English computer corpora: Selected papers and research guide*, ed. by Stig Johansson and Anna-Brita Stenström. 319-354. Berlin: Mouton de Gruyter.

Appendix 1: List of texts

Code	Author: Title orig./Title trans.
AB	Anita Brookner: *Latecomers/Etternølere*
AW	Alice Walker: *The Temple of My Familiar/Min følgesvenns tempel*
DF	Dick Francis: *Straight/Dødelig arv*
DL	Doris Lessing: *The Good Terrorist/Den gode terroristen*
FC	Finn Carling: *Under aftenhimmelen/Under the Evening Sky*
FW	Fay Weldon: *The Heart of the Country/Landets hjerte*
GS	Gunnar Staalesen: *I mørket er alle ulver grå/At Night All Wolves Are Grey*
HW	Herbjørg Wassmo: *Huset med den blinde glassveranda/The House With the Blind Glass Windows*
JA	Jeffrey Archer: *A Matter of Honour/En æressak*
JH	Joseph Heller: *Picture This/Se det*
KF	Knut Faldbakken: *Adams dagbok/Adam's Diary*
KFL	Kjartan Fløgstad: *Dalen Portland/Dollar Road*
LD	Len Deighton: *Berlin Game/Høyt spill i Berlin*
MA	Margaret Atwood: *Cat's Eye/Katteøyet*
RD	Roald Dahl: *Matilda/Matilda*
RR	Ruth Rendell: *Kissing the Gunner's Daughter/Brent barn*
SH	Stig Holmås: *Tordensønnen/Son-of-Thunder*
WG	William Golding: *The Paper Men/Papirmennene*