MULTILINGUAL CORPORA: MODELS, METHODS, USES

Stig Johansson*

ABSTRACT: This paper gives an account of our work on multilingual corpora at the University of Oslo. Different corpus models are presented, in particular the bi-directional translation model used in building the English-Norwegian Parallel Corpus (ENPC). The steps in building the ENPC are briefly outlined, with some comments on problems encountered. Most of the paper is concerned with the use of this sort of corpus. There is a brief introduction to the search program developed for the ENPC. The main point of the paper is to show that the bi-directional corpus model makes it possible to carry out contrastive studies and simultaneously control for translation effects. Bengt Altenberg's notion of mutual correspondence is introduced, with reference to his study of adverbial connectors in English and Swedish, based on the sister corpus of the ENPC, the English-Swedish Parallel Corpus. As an illustration of translation effects, there are some comments on the distribution of two verbs of posture in the ENPC: Norwegian *stå* and its English cognate *stand*. The difference is sharpest in original texts, while the distribution in the translations is clearly tinged by the source texts.

KEYWORDS: multilingual corpora, corpus models, corpus methodology, mutual correspondence, translation effects.

RESUMO: O presente artigo relata o trabalho com corpora multilíngües na Universidade de Oslo. Diferentes modelos de corpora são apresentados, em especial, o modelo bidirecional de tradução usado no English-Norwegian Parallel Corpus (ENPC). As etapas de construção do ENPC são de-

* University of Oslo, Norway.

scritas de forma sucinta, acompanhadas de alguns comentários acerca dos problemas encontrados. Grande parte do artigo é dedicada aos usos desse tipo de corpus. Há também uma breve introdução à ferramenta de busca desenvolvida para o ENPC. O principal objetivo aqui é mostrar que o modelo bidirecional de corpus de tradução possibilita a realização de estudos contrastivos e permite, simultaneamente, observar efeitos de tradução. A noção de correspondência mútua, introduzida por Bent Altenberg, é discutida quando comentamos seu estudo sobre conectores adverbias em inglês e sueco, desenvolvido com base no English-Swedish Parallel Corpus, criado segundo os mesmos parâmetros do ENPC. Como exemplo de efeitos de tradução, são tecidos alguns comentários quanto à distribuição de dois verbos de posição no ENPC: o norueguês stå e seu cognato em inglês stand. A diferença é mais marcada em textos originais; já nas traduções, a distribuição é claramente influenciada pelos textos originais.

UNITERMOS: corpora multilíngües; tipos de corpus; metodologia de corpus; correspondência mutual; efeitos de tradução.

1. Introduction

In the course of the last couple of decades there has been a rapidly increasing interest in corpus studies in linguistics, i.e. studies linked to text corpora. This is partly connected with the growing preoccupation among language researchers with the study of language in use, and partly it is related to the new possibilities of analysing large amounts of text using computers.

In this paper I am concerned with the development of multilingual corpora for use in contrastive analysis and translation studies. As an example, I will take our multilingual corpus project at the University of Oslo.

2. Models

The first step in our project was the development of the English-Norwegian Parallel Corpus (ENPC), a bi-directional translation corpus, with translations going both ways: English to Norwegian and Norwegian to English. Because it is structured in this way, we get a comparable corpus into the bargain. This is shown in Figures 1 and 2.

With a corpus of this kind we can make comparisons of different kinds, as shown by the arrows in Figure 3. Figure 4 shows what happens if we expand the model to three languages, as we have done at the University of Oslo in a project which we have undertaken in collaboration with the Department of Germanic Studies. We can compare:

- original texts in the three languages;
- original texts and translations across languages;
- original and translated texts in each language;
- translations across languages.

The main weakness of this model is that it is limited to texts that have actually been translated across the three languages.

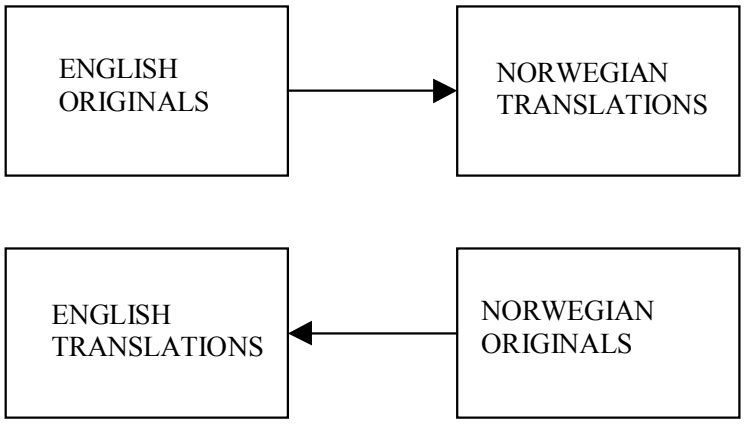

Figure 1: English and Norwegian: original texts and translations

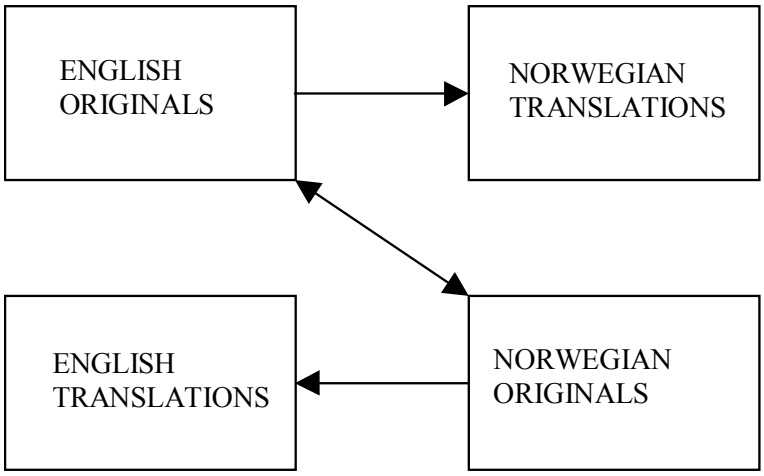

Figure 2: English and Norwegian: original texts in both languages

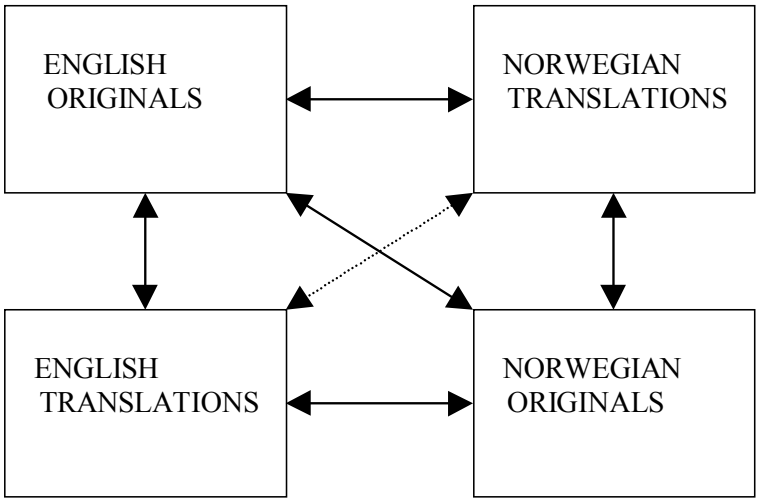

Figure 3: The model for the English-Norwegian Parallel Corpus

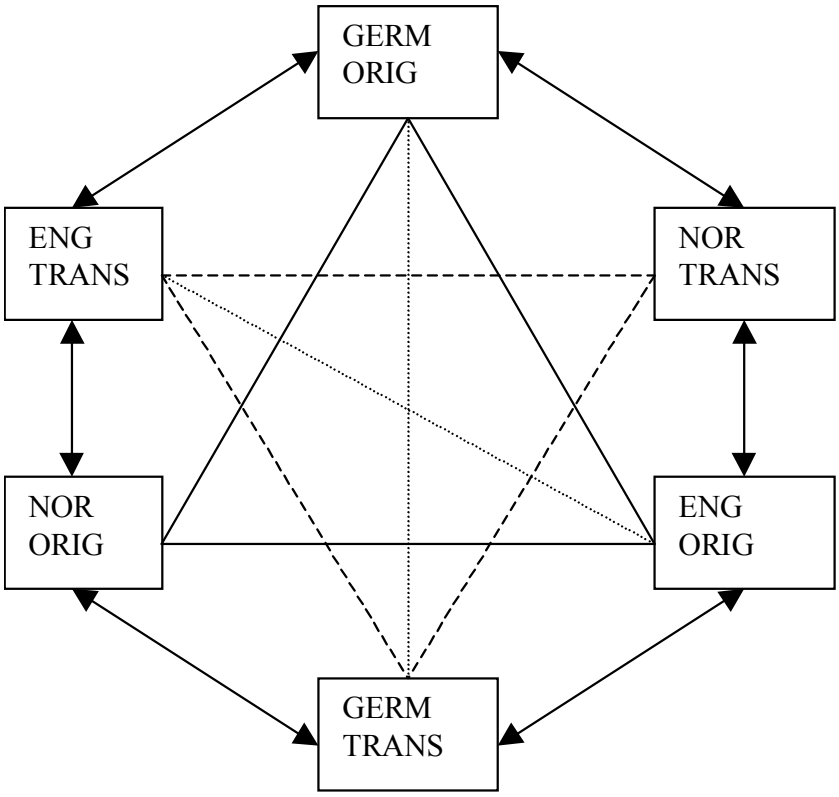

Figure 4: The Oslo Multilingual Corpus: English-Norwegian-German

Another model which we use is shown in Figure 5. At present we are building a corpus of Norwegian texts with their translations into English, German, and French, in cooperation with representatives from other language departments (German, French, and translation studies). With this corpus as well, we are restricted by the number of texts that have been translated into all of these languages. But we find it valuable to build this type of resource. The more languages we include, the more clearly can we see the characteristics of each language, and the more general questions can we ask about the nature of language and the characteristics of translation.

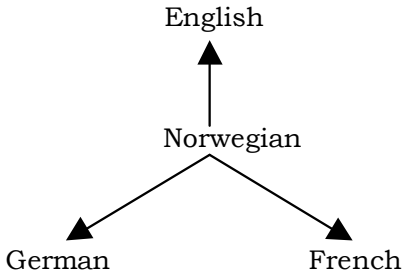

Figure 5: The Oslo Multilingual Corpus: Norwegian-English-German-French

One problem with most translation corpora is that there is just one translation for each text. To study the degree of variation in translation, we have compiled a small corpus according to the model shown in Figure 6. We have commissioned some of the best translators in Norway to translate two English texts that have not previously been translated into Norwegian. The translators have worked independently, and each has handed in both a draft and a final edited version.

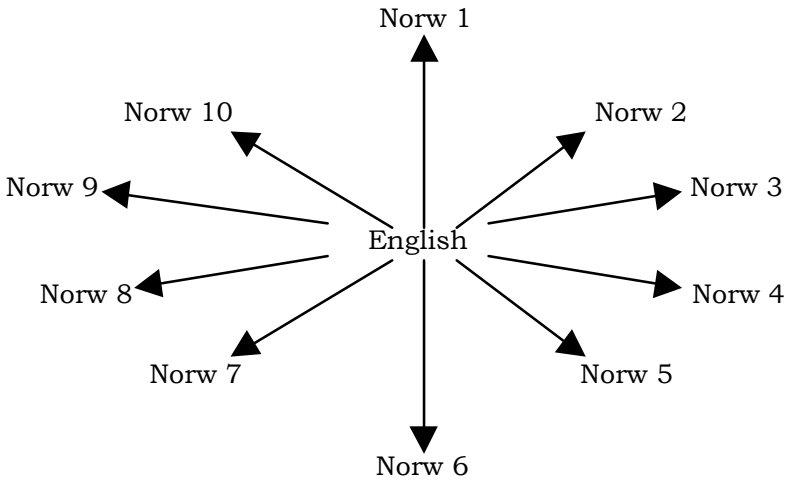

Figure 6: English and Norwegian: English source texts and multiple translations

These are the main models used in our multilingual corpus, which we now refer to as the Oslo Multilingual Corpus (OMC).

In addition to the languages I have mentioned, we also have some texts for English-Dutch, English-Portuguese and French-Norwegian. Since we have had cooperation with sister projects in Sweden and Finland, we also have the possibility of extending the comparison to Swedish and Finnish.

As already pointed out, translation corpora have some limitations (see also the point on text selection in Section 3 below). Hence, corpora of this kind must be supplemented by larger monolingual corpora in order to adequately represent the languages to be compared.

3. Methods

Space does not allow me to go into detail as regards our methodology. I will just mention the main steps in multilingual corpus building and briefly comment on some of them. For more details, see the English-Norwegian Parallel Corpus manual (www.hf.uio.no/iba/prosjekt).

• Text selection

To begin with, we make a survey of texts that have been translated between the languages we wish to include in the corpus. We focus on fairly recent texts, from the last 10-20 years or so, both fiction and non-fictional prose. The limitation to texts that have been translated means that we cannot hope to build corpora that could represent the languages involved in a fully satisfactory manner. The problem is made even more complicated by the fact that we want to build bidirectional corpora, where original texts in each of the languages are matched by genre and time of publication. The matching is difficult, as far more texts have been translated from the major European languages into Norwegian than in the other direction. As the corpus is expanded to include more languages, the problem becomes even more daunting.

To reduce the influence of idiosyncratic features, a consistent attempt is made to include a wide range of authors and translators. For the same reason, and also to reduce the problems in

getting permission from copyright holders (see the next point), we use text extracts rather than complete texts, in most cases extracts of 10,000 to 15,000 words. We try to match the material for each language, so that the different components of the corpus are comparable in size and extent. Thus the ENPC contains 50 original texts for each language, 30 fiction texts and 20 non-fictional texts, in all 200 texts including both originals and translations.

• Copyright clearance

One of the most difficult problems in building a corpus is getting copyright holders to grant permission to include texts . The problem is compounded by the fact that we must get permission both for the original texts and for the translations. A lot of correspondence is involved before we manage to get copyright clearance, and in many cases we never receive the permission we asked for, and selected texts must be discarded. The permission we get is quite restricted. The most important restrictions are that the texts can only be used for research and that the permission is limited to researchers at the University of Oslo and the University of Bergen. In our efforts to obtain copyright clearance, we have received valuable assistance from the authors' and translators' associations in Norway.

• Insertion of codes

The texts, both originals and translations, are scanned and then coded for a number of features, such as sentence (<s>), paragraph (<p>), and highlighting (<hi>), in accordance with the recommendations of the Text Encoding Initiative. The codes are inserted in the texts by means of a computer program.

• Proofreading and insertion of header

The texts are proofread both for scanning errors and coding errors. It is particularly important to check the coding of sen-

tences, or s-units, as the information on sentence division is crucial for the alignment stage that follows next. At this stage we also insert a header for each text, giving information both on the printed text and on the electronic version. The header coding is in accordance with the recommendations of the Text Encoding Initiative.

• Alignment

The most important stage is the alignment of originals and translations. This is done by a automatic alignment program developed by Knut Hofland: the Translation Corpus Aligner (see Hofland and Johansson, 1998). The program was originally developed for English-Norwegian, but has later been successfully adapted for many other language pairs. After alignment, each sentence has a unique identifier and a pointer (pointers) to the corresponding sentence(s) in the other language(s). The same program is used for all the models introduced in Section 2.

• Proofreading of the alignment

Although the alignment program has a high success rate, there are inevitable mistakes in the process. In the proofreading of the alignment we focus on sentences without a one-to-one sentence correspondence.

• Building of the database

After the alignment errors have been corrected, the texts are entered in a database in the format required by the search program developed for the project (see Section 4).

• Grammatical tagging

The English and Norwegian original texts have been grammatically tagged by means of a constraint grammar parser, in

collaboration with Atro Voutilainen, Helsinki, and the Text Laboratory at the University of Oslo. After tagging, each word form has a prefix that specifies the lemma and gives relevant grammatical information. For lack of resources, we have not been able to proofread and check the tagging in a systematic and exhaustive manner.

4. Uses

After the stages outlined above, the corpus is ready to be used. A special program has been developed for the corpus by Jarle Ebeling: the Translation Corpus Explorer (Ebeling, 1998). These are some features available in the program:

- It is possible to search for individual word forms or groups of word forms, e.g.: *take* or *take|takes|took|taking|taken*. Wildcards can be used, as in *take** for all words beginning with this character sequence.
- For grammatically tagged texts, it is possible to search for lemmas or word forms with particular tags, e.g. for all forms of the lemma *take*, tagged *<w l="take">*, or for the present tense form *takes*, tagged *<w p="Vpres">*.
- By using a filter we can limit the search to take into account only words in the surrounding context, e.g. *take* preceded within a specified span by the auxiliary *would* and/or followed within a specified span by the particle *up*.
- Perhaps the most important option from the point of view of translation studies is the possibility to specify what forms must or must not occur in the corresponding sentence in the other language(s). For an example, see below.
- The context of the search can be adjusted from single sentence pairs up to 25 sentences before or after the search item.
- It is possible to specify that the search item must be found in a particular position in relation to the beginning or the end of the sentence.

A couple of examples will suffice to illustrate the possibilities of the program.

In the search defined in Figure 7, we search for *heart* in ENPC/Fiction, in English original texts. The NOT filter at the bottom specifies that corresponding units in the Norwegian text must not contain an occurrence of *hjerte|hjertet*, i.e. the expected Norwegian translation. An example of a sentence found by this search is (1) below. The identity of the text is revealed by clicking on the code AT1.

(1) They were supposed to stay at the beach a week, but neither of them had the *heart* for it and they decided to come back early. (AT1)

De skulle egentlig vært på stranden en uke, men ingen av dem hadde *lyst* til å bli der lenger, så de bestemte seg for å dra hjem tidligere. [lit. 'had inclination to']

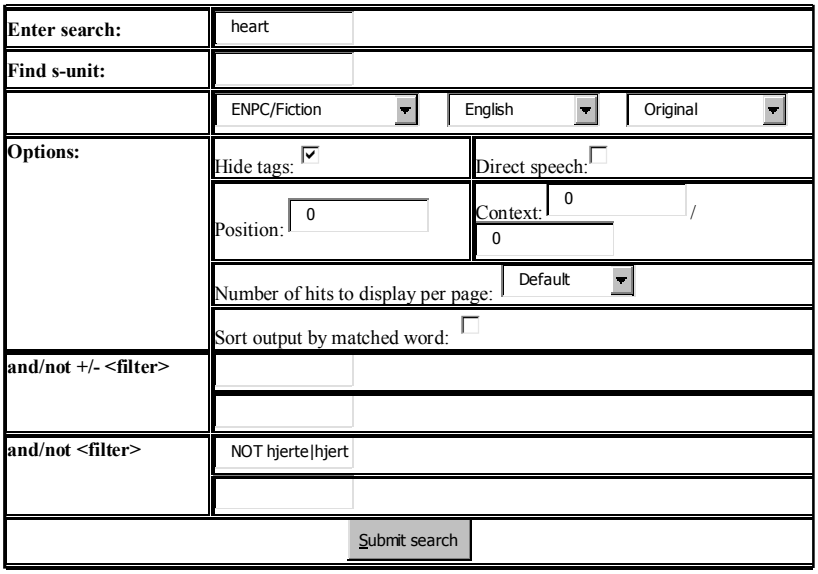

Figure 7: A search for *heart* using the Translation Corpus Explorer

There were 33 items of non-correspondence between *heart* and *hjerte|hjertet*, out of a total of 72 occurrences. Using the

AND filter instead, we find 39 instances. The search was done using the option 'Hide tags'. If we carry out the search with this option turned off, we get:

(2) <s id=AT1.1.s1 corresp=AT1T.1.s1>They were supposed to stay at the beach a week, but neither of them had the *heart* for it and they decided to come back early.</s>

<s id=AT1T.1.s1 corresp=AT1.1.s1>De skulle egentlig vært på stranden en uke, men ingen av dem hadde *lyst* til å bli der lenger, så de bestemte seg for å dra hjem tidligere.</s>

Here we see the coding that makes it possible to retrieve corresponding units from originals and translations (id= identifies the text and the number of the unit, corresp= identifies the corresponding unit in the other language). If we carry out the search in the tagged corpus, we get:

(3) <s id=AT1.1.s1 corresp=AT1T.1.s1><w p="Pnom">**They**</w> <w l="be" p="Vpast">**were**</w> <w l="suppose" p="EN">**supposed**</w> <w p="TO">**to**</w> <w p="Vinf">**stay**</w> <w p="PREP">**at**</w> <w p="DET">**the**</w> <w p="N">**beach**</w> <w p="DET">**a**</w> <w p="Nadv">**week**</w>, <w p="Cc">**but**</w> <w p="P">**neither**</w> <w p="PREP">**of**</w> <w l="they" p="Pobl">**them**</w> <w l="have" p="Vpast">**had**</w> <w p="DET">**the**</w> <w p="N">***heart***</w> <w p="PREP">**for**</w> <w p="Pobl">**it**</w> <w p="Cc">**and**</w> <w p="Pnom">**they**</w> <w l="decide" p="Vpast">**decided**</w> <w p="TO">**to**</w> <w p="Vinf">**come**</w> <w p="ADV">**back**</w> <w p="ADV">**early**</w>.</s>

<s id=AT1T.1.s1 corresp=AT1.1.s1><w p="Ppers">**De**</w> <w p="Vpretaux">**skulle**</w> <w p="ADV">**egentlig**</w> <w l="være" p="Vperfpaux">**vært**</w> <w p="PREP">**på**</w> <w l="strand" p="N">**stranden**</w> <w p="DETkvant">**en**</w> <w p="N">**uke**</w>, <w p="Cc">**men**</w> <w p="DETkvant"

p="DETkvant">**ingen**</w> <w p="PREP">**av**</w> <w l="de" p="Ppers">**dem**</w> <w l="ha" p="Vpretaux">**hadde**</w> <w p="N">***lyst***</w> <w p="PREP">**til**</w> <w p="Infmerke">**å**</w> <w p="Vinfaux">**bli**</w> <w p="ADV">**der**</w> <w l="lenge" p="Acmp" l="lang" p="Acmp">**lenger**</w>, <w p="Cc">**så**</w> <w p="Ppers">**de**</w> <w l="bestemme" p="Vpret">**bestemte**</w> <w p="Prefl">**seg**</w> <w p="PREP">**for**</w> <w p="Infmerke">**å**</w> <w p="Vinf">**dra**</w> <w p="ADV">**hjem**</w> <w l="tidlig" p="Acmp">**tidligere**</w>.</s>

To make it easier to read the text, I have given all the words in bold. Note that each word is accompanied by grammar information (p=) and, where applicable, also by lemma information (l=). Needless to say, this coding is not for the reader, but for use in specifying searches.

An example of another search is given in Figure 8. Here we look for occurrences of all words ending in *ing*, and with the grammatical tag ING, in the first position of sentences in the tagged English fiction texts.

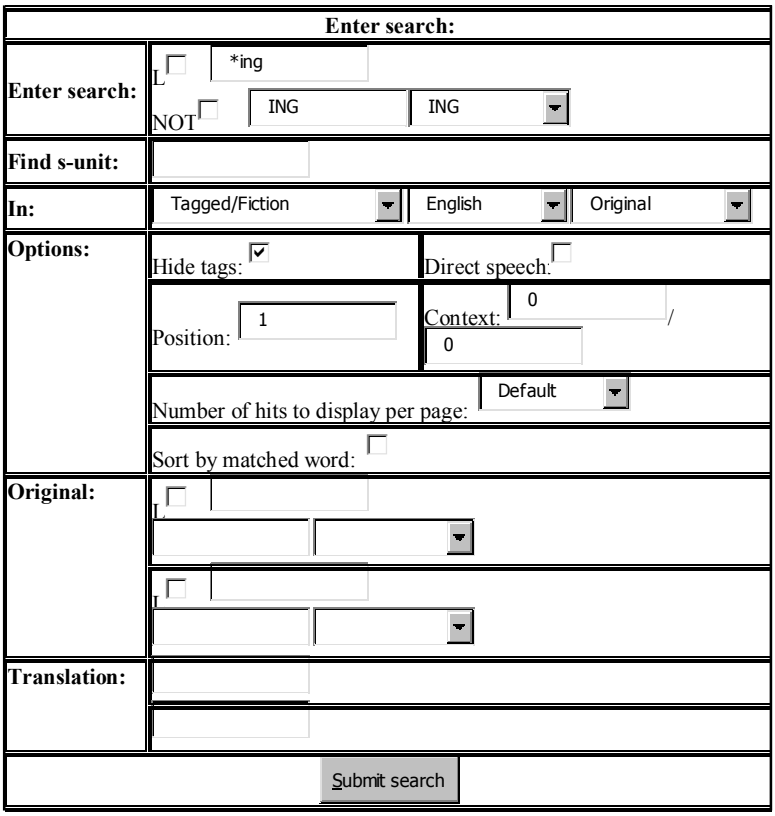

Figure 8: A search for sentences opening with an *ing*-form

An example from this search is:

(4) "*ING*">Lean*ing* up on his elbow, his face clenched with fury and disgust, he listened to what could be plainly heard even from his wall; then lay tense, "*ING*">breath*ing* fast. (DL2)

Han lente seg på albuen og lyttet, ansiktet var sammenbitt i raseri og avsky.
Så ble han liggende stiv. Han pustet tungt. [lit. 'He leaned himself on elbow-the and listened, face-the was together-bitten in rage and disgust. Then remained he lying stiff. He breathed heavily']

Note that both *ing*-forms here are highlighted, but the one we are interested in is the first one.

As a further illustration, I will include an example from a search for the Norwegian concessive marker *likevel* in our Norwegian-English-German-French corpus (cf. Figure 5), with the 'hide tags' option turned off:

(5) <s id=BHH1.3.3.s400 corresp='BHH1TE.3.3.s402 BHH1TF.3.3.s416 BHH1TD.3.3.s308'>Det er to år og to måneder til jeg har examen artium, hvis jeg ikke har gitt tapt og falt av lasset lenge før den tid.</s><p id=BHH1.3.3.p107> <s id=BHH1.3.3.s401 corresp='BHH1TE.3.3.s403 BHH1TF.3.3.s417 BHH1TD.3.3.s309'>Kanskje er det *likevel* en feil ved meg at jeg er så tålmodig.</s><s id=BHH1.3.3.s402 corresp='BHH1TE.3.3.s404 BHH1TF.3.3.s418 BHH1TD.3.3.s310'>Jeg samler opp og samler opp, og nåde meg den dagen jeg har fått nok og det brister for meg.</s><p id=BHH1.3.3.p108> (BHH1)

<s id=BHH1TD.3.3.s308 corresp='BHH1.3.3.s400 BHH1TE.3.3.s402 BHH1TF.3.3.s416'>Erst in zwei Jahren und zwei Monaten werde ich mein Abitur haben, wenn ich nicht längst vorher aufgegeben habe!</s><p id=BHH1TD.3.3.p88> <s id=BHH1TD.3.3.s309 corresp='BHH1.3.3.s401 BHH1TE.3.3.s403 BHH1TF.3.3.s417'>Vielleicht ist meine Geduld *ja doch* ein Fehler.</s><s id=BHH1TD.3.3.s310 corresp='BHH1.3.3.s402 BHH1TE.3.3.s404 BHH1TF.3.3.s418'>Ich sammele und sammele alles in mir, und Gnade mir an dem Tag, an dem ich genug habe und alles aus mir herausbricht.</s> (BHH1TD)

<s id=BHH1TE.3.3.s402 corresp='BHH1.3.3.s400 BHH1TF.3.3.s416 BHH1TD.3.3.s308'>It will be two years and two months until I have my artium degree, if I don't give up and drop by the wayside before that time.</s><p id=BHH1TE.3.3.p110> <s id=BHH1TE.3.3.s403 corresp='BHH1.3.3.s401 BHH1TF.3.3.s417

BHH1TD.3.3.s309'>Maybe it is a defect, *after all*, that I am so patient.</s><s id=BHH1TE.3.3.s404 corresp='BHH1.3.3.s402 BHH1TF.3.3.s418 BHH1TD.3.3.s310'>I keep holding things in, and God help me the day I've had enough and something snaps.</s><p id=BHH1TE.3.3.p111> (BHH1TE)

<s id=BHH1TF.3.3.s416 corresp='BHH1.3.3.s400 BHH1TE.3.3.s402 BHH1TD.3.3.s308'>J'en ai encore pour deux ans et deux mois avant de passer le baccalauréat à moins que, pris de découragement, je ne décide de tout lâcher d'ici là.</s><p id=BHH1TF.3.3.p108> <s id=BHH1TF.3.3.s417 corresp='BHH1.3.3.s401 BHH1TE.3.3.s403 BHH1TD.3.3.s309'>*Finalement*, peut-être ai-je tort de me montrer si patient.</s><s id=BHH1TF.3.3.s418 corresp='BHH1.3.3.s402 BHH1TE.3.3.s404 BHH1TD.3.3.s310'>J'accumule, j'accumule, jusqu'au jour où la mesure sera comble et je ne tiendrai plus.</s><p id=BHH1TF.3.3.p109> (BHH1TF)

Note the pointers between all the versions. These make it possible to define searches in each language and retrieve the corresponding units in all the languages included.

The Translation Corpus Explorer is a search program. That is, if we want to further the analysis of the data by computer, we need to import the results of searches into some other program. The search program provides material that is used in studies of different kinds. Some examples of studies we have carried out so far are (for references, see our websites):

- presentative constructions in English and Norwegian (Ebeling)
- word order in English and Norwegian (Hasselgård)
- expressing possibility in English and Norwegian (Løken)
- Norwegian discourse particles in a contrastive perspective (Johansson and Løken)
- the Norwegian concessive marker *likevel* and its correspondences in English (Fretheim and Johansson)

- the construction type *that's what* and corresponding expressions in Norwegian and German (Johansson)
- generic subjects in English, Norwegian, and German (Johansson)
- relative constructions in French and Norwegian (Helland)
- English *ing*-constructions and their correspondences in Norwegian (Behrens)
- clauses introduced by the conjunction *indem* (German) and corresponding expressions in English and Norwegian (Behrens and Fabricius-Hansen)
- information density in Norwegian and German (Fabricius-Hansen, Solfjeld)
- explicitation in translation (Øverås)
- deviant features in translations compared with original texts (Elsness, Hasselgård, Johansson)

See also the collections of papers edited by Aijmer *et al.* (1996), Johansson and Oksefjell (1998), Borin (2002), and Hasselgård *et al.* (2002).

The most important point to note is that our multilingual corpora provide a good basis both for contrastive analysis and for translation studies. In other words, we can use the corpora both to gain insight into languages and to reveal translation effects. I will try to show this by a couple of examples.

5. Studying language through translation

Long before the age of computer corpora, in a paper on 'the translation paradigm', Levenston suggested that contrastive statements

> … may be derived from either (a) a bilingual's use of himself as his own informant for both languages, or (b) close comparison of a specific text with its translation. (Levenston, 1965: 225)

The use of multilingual corpora, with a variety of texts and a range of translators represented, increases the validity and re-

liability of the comparison. It can be regarded as the systematic exploitation of the intuition of translators, as it is reflected in the pairing of source and target language expressions in the corpus texts.

As an example of a corpus-based contrastive investigation, consider Bengt Altenberg's study of adverbial connectors in English and Swedish, based on the sister corpus of the ENPC, the English-Swedish Parallel Corpus. Altenberg introduces the notion of *mutual correspondence* (MC), which is calculated 'by means of the simple formula

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

where A_t and B_t are the compared categories or items in the translations, and A_s and B_s the compared categories or items in the source texts. The value will range from 0% (no correspondence) to 100% (full correspondence).' (Altenberg, 1999:254)

To take an example from Altenberg's study: English *instead* (30 occurrences, 24 of which were rendered by *i stället*, i.e. 80%) and Swedish *i stället* (41 occurrences, 32 of which were rendered by *instead*, i.e. 78%). The calculation of mutual correspondence is as follows:

$$\frac{(24 + 32) \times 100}{30 + 41} = 79$$

The mutual correspondence value in this case is quite high, approximately 79%.

Using this measure Altenberg can reveal the mutual correspondence both of different types of adverbial connectors (highest with listing connectors, such as *to begin with* and *to conclude*, and lowest with transitional and explanatory conjuncts, such as *incidentally* and *after all*) and of individual connectors (see Table 1, quoted from Altenberg, 1999: 257).

Table 1: Mutual correspondence of individual connectors (n ≥ 10 in source texts)

English ↔	Swedish	MC %	Translation bias %		Omiss.
			E → S	S → E	%
instead	*i stället*	79	80	78	18
on the other hand	*å andra sidan*	74	74	75	5
then (inferential)	*då* (inferential)	65	80	56	39
however	*emellertid*	63	47	81	8
that is to say (i.e.)	*det vill säga (dvs)*	55	57	54	48
therefore	*därför*	54	62	49	10
...	...	...	...	...	...
so	*alltså*	15	11	24	22
nevertheless	*dock*	13	0	15	0
anyway	*i varje fall*	12	6	20	18
therefore	*alltså*	10	18	3	24
thus	*sålunda*	4	0	18	27
say	*till exempel (t ex)*	4	50	1	8
–	*nämligen* (explan.)	0	–	–	47
now (transitional)	–	0	–	–	84

The table shows that mutual correspondence values vary widely. There are no cases of 100% correspondence, though the pairs at the top of the table are very often used to translate each other. For the pairs at the bottom, the degree of intertranslatability is low, either because 'there is a better choice in the other language or [because] there is a lexical gap in the conjunct system' (Altenberg, *ibid.)*, as with the English transitional conjunct *now* and the Swedish explanatory conjunct *nämligen*. The intertranslatability sometimes varies strikingly depending upon the direction of translation, indicated by *translation bias* in Table 1. For example, four out of five instances (81%) of the Swedish connector *emellertid* are translated by *however*, which in turn is rendered by *emellertid* in less than half (47%) of the cases.

Another point shown in the table is the degree to which individual connectors are omitted in the translation. The rate of omission is naturally very high where there is a lexical gap, but it is sometimes high in other cases as well. In a recent paper, coauthored with Karin Aijmer (Aijmer and Altenberg, 2002), there is a detailed discussion of such *zero correspondences* and the grounds for omission.

Altenberg goes on to set up paradigms of correspondences both for individual connectors (e.g. the various translations of

English *so*) and for a whole set of connectors, showing what forms are available in each language and their intertranslatability across the languages. Last but not least, he illustrates and discusses the use of connectors in context by making use of the rich material in the corpus. This study is an exemplary one, which not only brings new insight into the study of connectors, but also provides a model for corpus-based contrastive studies in general.

6. Studying translation through corpora

There is no doubt that translations provide a good means of studying relationships between languages, and they also serve to bring out features of the individual languages that might be difficult to see otherwise. But as translations represent a special use of the language that may differ in important ways from the language of original texts in the target language, it is important to control for translation effects. This control function is built into corpora structured according to the models shown in Figures 3 and 4 above. As an illustration, I will briefly examine two verbs of posture: Norwegian *stå* and its English cognate *stand*. Figure 9 shows the distribution of the two verbs in the fiction texts of the ENPC.

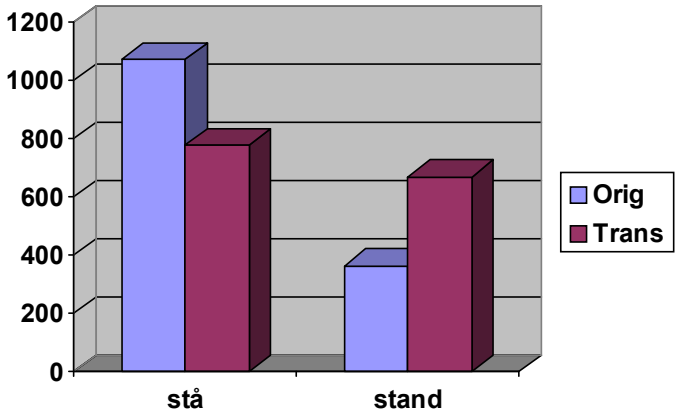

Figure 9: The frequency of *stå* and *stand* in the fiction texts of the ENPC

As the material for the two languages is approximately equal in size, we can compare raw frequencies. We see that there is a vast difference between the frequency of the two verbs in original texts, which indicates that the Norwegian verb has a wider range of use. Though absolute frequencies differ, we find the same relationship for *ligge* vs. *lie* and *sitte* vs. *sit*. The three verb pairs are also comparable as regards the frequencies in translated vs. original texts: the frequencies go down for *stå/ligge/sitte* and up for *stand/lie/sit*. In other words, frequency differences are evened out in translated text. This is a pattern we have frequently observed in our corpus studies (see e.g. Johansson, 2001).

Inspection of the corpus material shows that *stå* has a wide range of correspondences in English, including a large number of verbs other than its cognate *stand* (often *be*, but not infrequently semantically richer verbs). There is also a good deal of zero correspondence. It is uncertain precisely how the relationship between the English and the Norwegian verb should be characterised, as both have a number of senses and take part in many more or less fixed sequences. A possible generalisation is that the Norwegian verb often has a weaker meaning than its English counterpart and approaches a mere copula (cf. *estar* in Spanish and Portuguese, derived from a verb meaning 'stand'). Alternatively, we might ascribe the difference to a stronger tendency in Norwegian to focus on the type of posture, where English is content to indicate position. But an in-depth contrastive study goes far beyond the scope of this paper so I will merely make brief comments on the apparent overuse of the English verb *stand* in texts translated from Norwegian.

Cases where the use of the English verb is marked, or even unacceptable, are not difficult to find. Some examples are:

(6) I garasjen *står* fem par terrengski, tre par slalåmski, en scooter, sykler, hageredskap m.m. (BV2)
In the garage, *stand* five pairs of cross-country skis, three pairs of slalom skis, a scooter, bicycles, garden equipment etc.

(7) Hun ser sløvt på trærne langs veien, bladverket er nytt og lyst, i hagene *står* tulipaner, gresset er klipt alt. (BV2)

She looks listlessly at the trees along the road; the foliage is new and fresh, in the gardens *stand* tulips, the grass has already been cut.

(8) Nederst i bokhyllen *lå* det noen gamle ukeblad, og over dem *sto* det en håndfull bøker, sannsynligvis fra hennes barndom. (FC1)
Some old weekly magazines *were lying* at the bottom of the bookcase, and above them *stood* a handful of books, probably from her childhood.

(9) På en hylle *sto* små keramikkvaser sammen med stener og potteskår, helt verdiløst, men på en merkelig måte fornemt allikevel. (KF1)
On a shelf *stood* two ceramic vases, together with stones and potsherds, of no value whatsoever but strangely refined anyway. (KF1T)

(10) Den svarte vedkomfyren *sto* ikke i kroken der den alltid hadde *stått*. (KF2)
The black wood-burning stove *was* not in the corner where it had always *stood* before.

In example (8) we note two verbs of posture, which have both been translated by their cognate English verbs, in (10) there are two forms of *stå*, one of which has been rendered by *was* and the other one by *stood*.

The Norwegian verb *stå* seems to pose a problem to English translators. On the one hand, there is the danger of overuse of the English cognate. On the other hand, there is the choice between a wide variety of verbs, ranging from *be* to semantically much richer verbs. One reflection of the dilemma is that translators who opt for the cognate verb often insert the adverbial *there*, although there is no equivalent adverbial in the Norwegian source text, as in:

(11) I stuen *står* fru Schøning, hun *står* i rød genser og sorte fløyelsbukser og spiller fiolin. (BV2)
Mrs Schøning *is* in the living room, she *stands there* in a red sweater and black velvet trousers playing the violin.

(12) Som frosset *sto* jeg i mørket. (MN1)
As if frozen, I *stood there* in the dark.

The combination *stand + there* is, in fact, about eight times commoner in the translations from Norwegian than in the English original texts (again in an approximately equal amount of text).

7. Concluding remarks

I hope my examples have shown that a well-constructed multilingual corpus can be used to advantage both for contrastive analysis and translation studies. A contrastive study becomes more than an abstract comparison of language systems; systems *are* revealed, but they are connected with the use of languages in context. This kind of study is capable of uncovering translation characteristics which, besides giving insight into particular translation problems, may help us understand the nature of translation. There are important applications for language teaching, bilingual lexicography and the training of translators.

A corpus including only translations in one direction produces results that may be difficult to interpret. The special advantage of the bidirectional translation model is that it makes it possible to distinguish between language differences and translation effects. This is the main message I have tried to convey.

References

AIJMER, K.; ALTENBERG, B. and JOHANSSON, M. (eds.) (1996) *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994*.
Lund Studies in English 88. Lund, Lund University Press.

AIJMER, K. and ALTENBERG, B. (2002) Zero translations and cross-linguistic equivalence: evidence from the English-Swedish Parallel Corpus. In: Breivik, L.E. and Hasselgren, A. (eds.) *From the COLT's mouth … and others. Language corpora studies on honour of Anna-Brita Stenström*. Amsterdam & New York, Rodopi, p. 19-41.

ALTENBERG, B. (1999) Adverbial connectors in English and Swedish: semantic and lexical correspondences. In: Hasselgård, H. and Oksefjell,

S. (eds.), *Out of corpora. Studies in honour of Stig Johansson*. Amsterdam, Rodopi, p. 249-68.

BORIN, L. (ed.). (2002) *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April, 1999*. Amsterdam & New York, Rodopi.

EBELING, J. (1998) The translation corpus explorer: a browser for parallel texts. In: Johansson, S. and Oksefjell, S. (1998), p. 101-12.

HASSELGÅRD, H.; JOHANSSON, S.; BEHRENS, B. and FABRICIUS-HANSEN, C. (eds) (2002) *Information structure in a cross-linguistic perspective.* Amsterdam & Atlanta, GA, Rodopi.

HOFLAND, K. and JOHANSSON, S. (1998) The Translation Corpus Aligner: a program for automatic alignment off parallel texts. In: Johansson, S. and Oksefjell, S. (1998), p. 87-100.

JOHANSSON, S. (1998) On the role of corpora in cross-linguistic research. In: Johansson, S. and Oksefjell, S. (1998), p. 3-24.

JOHANSSON, S. (2001) Translationese: evidence from the English-Norwegian Parallel Corpus. In: Allén, S.; Berg, S.; Malmgren, S.G.R.; Norén, K. and Ralph, B. (eds.) *Gäller stam, suffix och ord. Festskrift till Martin Gellerstam den 15 oktober 2001*, p. 162-76. Göteborg, Elanders Novum.

JOHANSSON, S. and HOFLAND, K. (1994) Towards an English-Norwegian parallel corpus. In: Fries, U.; Tottie, G. and Schneider, P. (eds.) *Creating and using English language corpora*. Amsterdam & Atlanta, GA, Rodopi, p. 25-37.

JOHANSSON, S. and OKSEFJELL, S. (eds.) (1998) *Corpora and cross-linguistic research: theory, method, and case studies*. Amsterdam & Atlanta, GA, Rodopi.

LEVENSTON, E.A. (1965) The 'translation paradigm': a technique for contrastive syntax. *International Review of Applied Linguistics* 3, p. 221-5.

Web sites

The English-Norwegian Parallel Corpus: www.hf.uio.no/iba/prosjekt
The Oslo Multilingual Corpus: www.hf.uio.no/german/sprik