SOME WAYS OF LOOKING INTO A MULTILINGUAL CORPUS

(PREPARED FOR THE 19TH EUROPEAN SYSTEMIC FUNCTIONAL LINGUISTICS CONFERENCE AND WORKSHOP, 23RD - 25TH JULY 2007, SAARBRÜCKEN, GERMANY)

Stig Johansson, University of Oslo

1 INTRODUCTION

Thank you very much for the invitation to give a plenary lecture on this occasion. I have learned a great deal from systemic functional linguistics, not least about the importance of studying language in use. To observe language in use it is crucial to study evidence from corpora. By corpora I understand collections of texts that have been put together in a principled way for the study of language. The use of corpora in electronic form makes it possible to reveal patterns which may be hard to see by ordinary observation. Great advances have been made. To mention just two major contributions, we have Douglas Biber's work on language variation and John Sinclair's studies of lexis and collocational patterns.

As John Sinclair recently passed away, I would like say a few words about him and his work. He realised early that "[…] if one wishes to study the 'formal' aspects of vocabulary organization, all sorts of problems lie ahead, problems which are not likely to yield to anything less imposing than a very large computer" (Sinclair 1966: 410). Later in the paper we read that "it is likely that a very large computer will be strained to the utmost to cope with the data" (p. 428). There was no way of knowing what technological developments lay ahead, and that we would get small computers with an infinitely larger capacity than the large computers at the time this was written.

Around this time John Sinclair compiled the world's first electronic corpus of spoken language. The corpus was fairly small, about 135,000 words, but considering the difficulties of recording, transcribing, and computerising spoken material, this was quite an achievement. We can read about the project in a book published a couple of years ago (Sinclair et al. 2004). The book is significant both because it gives access to the OSTI Report, which had been difficult to get hold of, and because of the interview, which gives insight into the development of John Sinclair's thinking. In the interview John Sinclair says that he did very little work on corpora in the 1970s, frustrated by the laboriusness of using the corpus and by the poor analysis programs which were available. But he and his team at Birmingham did ground-breaking work on discourse, leading to an important publication on the English used by teachers and pupils (Sinclair and Coulthard 1975). As I have understood it, what was foremost for John Sinclair was his concern with discourse and with studying discourse on the basis of genuine data.

We must "trust the text", as he put it in the title of a recent book. This applies both to the discourse analysis project and to his corpus work.

The 1980s was the great breakthrough for the use of corpora in lexical studies. John Sinclair and his team in Birmingham started the building of a large corpus and initiated the COBUILD project which led to the first corpus-based dictionary: *The Collins COBUILD English Language Dictionary* (Sinclair et al. 1987). There were a number of innovative features of this dictionary. Later dictionaries have not followed suit in all respects, but it is to the credit of John Sinclair and his team that English dictionaries these days cannot do without corpora.

Later he developed his ideas in a steady stream of conference papers, articles, and books. He was a multifaceted man. He was concerned both with linguistic theory and its applications, above all in lexicography. There is a remarkable consistency, all the way back to his paper "On beginning the study of lexis" (Sinclair 1966). By consistency I do not mean stagnation. What was consistent was his way of thinking – original, always developing, yet never letting go of the thought that the proper concern of linguistics is to study how language is actually used and how it functions in communication – through corpora, via lexis, to discourse.

Those who work with corpora have often been misunderstood as being mere data gatherers. But the data must be interpreted. John Sinclair taught us how we can see new things by systematically collecting texts in electronic form, examining them using new computer tools, and carefully interpreting the evidence. His way of working can serve as a model for linguists searching for new insight and as a source of inspiration at a conference on data and interpretation in linguistic analysis.

2 THE IMPORTANCE OF MULTILINGUAL CORPORA

I turn now to the main theme of my talk. In recent years there has been a fast increasing interest in multilingual corpora. What is the relevance of multilingual corpora in linguistic research? As I see it, they are important for a number of reasons:

- They make it possible to compare languages in a systematic manner, including preferences in language use.
- The comparison throws the characteristics of the individual languages into relief and gives evidence of typological as well as of universal features.
- Multilingual corpora consisting of original texts and their translations provide a way of making meaning visible.
- They can be used to reveal characteristics of translated vs. original texts.
- They allow applications in foreign-language teaching, lexicography, translator training, and language engineering.

Not least, "they give new insights into the languages compared – insights that are likely to be unnoticed in studies of monolingual corpora" (Aijmer and Altenberg 1996: 12).

One of the centres for multilingual corpus work is here in Saarbrücken. We have the monographs by Elke Teich (2003), Silvia Hansen (2003), and Stella Neumann

(2003). Right now Erich Steiner and his team are engaged in a very interesting project developing a multi-register corpus for contrastive analysis and translation studies (Steiner 2005). In my talk I will report on work we have done at the University of Oslo. The focus will be on ways of looking into a multilingual corpus (for more detail, see Johansson 1998 and 2007).

3 TYPES OF MULTILINGUAL CORPORA

Before we talk about ways of looking into a multilingual corpus, we need to define what is meant by a multilingual corpus. By this I mean collections of texts in two or more languages that are matched in some way, either because they are in a translation relationship or because they share characteristics with respect to genre, time of production, etc. We can call these two types ***translation corpora*** vs. ***comparable corpora***. Both have their advantages and their limitations, but they can be combined within the same framework, as we have done with our English-Norwegian Parallel Corpus (ENPC). The original corpus model is shown in Figure 1.

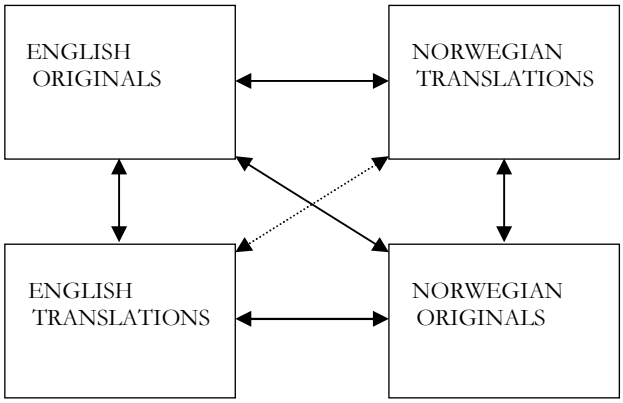

Figure 1 The model for the English-Norwegian Parallel Corpus

With the ENPC we can compare original texts across the two languages as well as original texts and their translations. The double arrows indicate that the comparison can start in either direction. We can also compare original vs. translated texts within each language to reveal possible translation effects, and we can study translated texts in the two languages to reveal general characteristics of translation. So the corpus changes depending upon our point of view and allows different ways of seeing.

When the model is expanded to three languages, the picture gets a bit more complicated. Figure 2 shows how we represent our English-German-Norwegian corpus; we call it the ***diamond model***. This allows the same types of comparison as the ENPC. The main problem with bidirectional corpora of this kind is that we are

limited to the types of texts that are translated and that it may be difficult to match texts across languages.

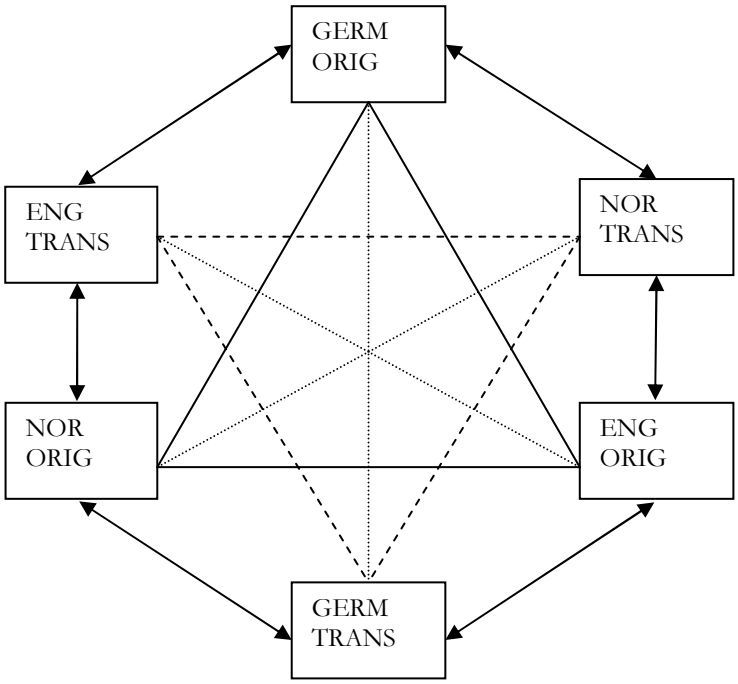

Figure 2 The Oslo Multilingual Corpus: English-Norwegian-German

In addition to the models I have shown, we have a number of other subcorpora under the general umbrella of the Oslo Multilingual Corpus. I will come back to this later. To start with I would like to introduce the notion of correspondence.

4 CORRESPONDENCE TYPES

With a corpus like the ENPC we can observe ***correspondences***. I will illustrate this with reference to the Norwegian modal particle *nok*. Etymologically, this is related to German *genug* and English *enough*, and it can also be used in the same way as these words. In addition, it can be a modal particle, which can be roughly described as indicating the speaker/writer's assessment of the probability of a situation. In the corpus we see that the particle *nok* has a large number of ***overt*** correspondences in English, chiefly adverbs (*probably, undoubtedly*, etc.), verb forms (*must, be bound to*, etc.), and comment clauses (*I suppose, I think*, etc.). In other words, there is a high proportion

of ***divergent*** correspondences, with forms belonging to different categories in the two languages. Some examples are:[1]

(1) Det er *nok* ål i gratengen likevel. (LSC1)
So there *probably* is eel in the soufflé.

(2) Etterpå gråter en av guttene, det er *nok* William. (BV2)
Afterwards one of the boys starts to cry, it *must* be William.

(3) Så det nærmer seg *nok* slutten. (KA1)
So *I suppose* the end is near.

Table 1 lists the main correspondences, both ***translations*** and ***sources***, i.e. forms in the English source texts which correspond to the particle in the Norwegian translations.

Table 1 Correspondences of the Norwegian modal particle *nok*, expressed in per cent within each column

Correspondences	Norw orig Eng transl	Norw transl Eng orig
probably	25	6
other adverb	21	4
verb construction	11	10
clause	9	10
miscellaneous	3	5
zero	31	65
Total (raw freq)	141	79

The most striking finding is the high frequency of ***zero correspondence***, i.e. instances where the English text does not contain any form that can be related specifically to the Norwegian modal particle. The frequency is particularly high in the case of English sources: two thirds of the instances of *nok* in the Norwegian translations appear to come from nowhere. It is notable that the next most frequent English sources are verb constructions and clauses, indicating that these are perceived by Norwegian translators to be more similar to the Norwegian particle than English adverbs.

What happens if we reverse the perspective and examine Norwegian correspondences of *probably*, i.e. the translation most frequently found for Norwegian *nok*? The results are given in Table 2. We notice, first of all, that the frequency of zero correspondence is low, regardless of the direction of translation. The great majority of the translations are ***congruent***, i.e. adverbs like the English counterpart. The most common translations of *probably* (*sannsynligvis* and *antageligvis*) match the English form

[1] The original version is generally listed first and is accompanied by a reference code. For an explanation of the reference codes, see: http://www.hf.uio.no/iba/prosjekt/index.html

both with respect to grammar and meaning. *Nok* and another Norwegian particle, *vel*, are rarely used to render *probably*, although the two together are the source of more than half of the instances of *probably* in the English translated texts.

Table 2 Correspondences of the English adverb *probably*, expressed in per cent within each column

Correspondences	Eng orig Norw transl	Eng transl Norw orig
nok	3	25
vel	6	28
antagelig(vis)	21	3
kanskje	3	9
sannsynligvis	37	16
sikkert	11	9
trolig	3	1
miscellaneous	13	6
zero	2	4
Total (raw freq)	94	141

A plausible interpretation of these results is that the existence of close formal and semantic correspondences of *probably* simplifies the task of the Norwegian translator, who can stay close to the original text. In contrast, when faced with the problems of rendering Norwegian *nok*, the English translator finds no easy solution. Most typically the meaning is either left unexpressed (zero correspondence) or strengthened (by the use of *probably* or some other adverb).

Both in the case of *nok* and *probably* we find a marked difference in distribution between original and translated texts, with underuse in the translations in one case, and overuse in the other: *nok* was underused (79 instances in the translations vs. 141 instances in the original texts), and *probably* overused (141 instances in the translations vs. 94 in the original texts).[2] The lack of a clear counterpart in the English source leads to underuse of *nok* in the Norwegian translations. *Probably*, on the other hand, has close counterparts in Norwegian, and it seems to have been pushed beyond its ordinary use by the translator's attempt to render the Norwegian modal particles *nok* and *vel*.

Although at the outset *nok* and *probably* would seem to be quite similar, in that they both indicate the speaker/writer's assessment of the probability of a situation, the correspondence patterns show that they are rather different in use. To summarise so far, we can classify cross-linguistic correspondences according to whether they are translations or sources, overt or zero, and syntactically congruent or divergent; see Figure 3. For example: *nok* > *I suppose* is a translation, overt, and divergent.

[2] As the subcorpora of original and translated texts are equal in size (number of texts and approximate number of words), we can compare raw frequencies.

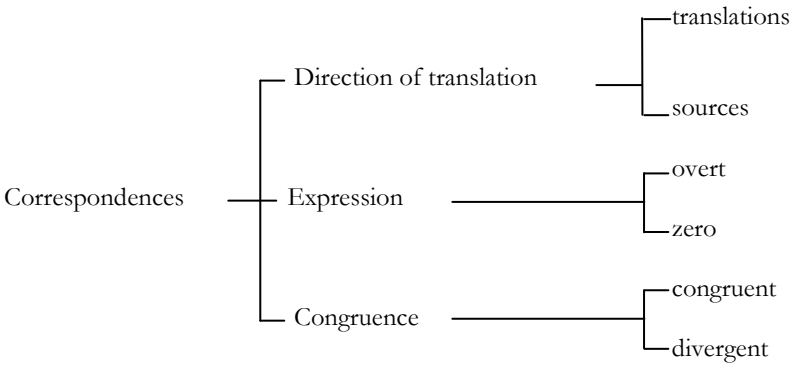

Figure 3 Classification of correspondences

Correspondences are what we can observe, but they need to be interpreted. Divergent correspondences commonly occur even in the case of closely related languages such as English and Norwegian. They can be taken to indicate to what extent the repertoire of forms used for particular purposes differs across languages. Zero correspondence is often found where there is no natural match across languages, and particularly in the case of forms expressing interpersonal and textual (rather than ideational) meaning. Zero correspondence goes both ways and applies both to forms in the source text which have no formal counterpart in the target text and to forms in the target text which seem to have appeared out of the blue, although there is no formal counterpart in the source text. We can speak of zero correspondence by ***omission*** vs. ***addition***.

Where there is omission, there may be compensation in the linguistic context, i.e. the meaning may be (partially) carried by some other form. Alternatively, the meaning has to be inferred, or it may be lost altogether. Addition can be interpreted as the translator's response to the whole context, reflecting cross-linguistic differences in the sorts of meanings that are conventionally expressed in natural discourse. I will now illustrate this with reference to a familiar English word: the discourse particle *well* and its correspondences in Norwegian and German.

5 TRANSLATION PARADIGMS

The term translation paradigm is generally used with reference to different approaches to translation, but here I mean the set of forms in the target text which are found to correspond to particular words or constructions in the source text; or the other way around: the set of forms in the source text which are found to correspond to particular words or constructions in the target text. Let's look at the translation paradigm we get for English *well* in translation into Norwegian and German; see Table 3.[3] At the time of the investigation, the English-German-Norwegian subcorpus contained 32 English

[3] For more detail, see Johansson (2006).

original texts – with a couple of exceptions they were all fiction texts – with translations into both Norwegian and German. Of these, 26 contained examples of the discourse particle *well*, in all 226 instances.

Table 3 *Well* in English original texts: German and Norwegian translations

Form type	German		Norwegian	
Discourse particles (DP)	**also**	**38**	altså	1
	also gut	3		
	also hör mal	1		
	also schön	1		
	ja	6	**ja**	**24**
	ja also	1	ja ja (ja, ja)	4
			ja … i hvert fall ('in any case')	1
			jaså	2
			javel	3
			jo	3
			jo… kan du skjønne ('you see')	1
			nei ('no')	6
			nei … men ('but')	1
			nei, nå får det være nok ('enough!')	1
	na	**12**	nå	7
	na ja (naja)	**24**	**nåja (nå ja)**	**11**
	na gut	1	nåvel (nå vel)	7
	na schön	6		
	na und?	1		
	nun	**17**		
	nun ja	**17**		
	nun gut	1		
	nun sag mal	1	**tja**	**13**
	tja	**11**	tja, sannelig ('truly')	1
			vel	**57**
			vel ... i hvert fall ('in any case')	1
Modal particles (MP)	eben	1	da	7
	ja	3	jo	3
DP + MP	na … ja	2	ja ... jo ... da	1
	na ja … doch	2	javel, greit da ('OK then')	1
	na ja … eben	1	ja ... jo	2
	na ja … schon	2	nei ... jo	2
	nun ja … eben	2	vel … da	1
	tja … eben	1		
Conjunctions	aber	7	**men**	**10**
			men … nå (MP)	1
	oder	1		
	und	3	og … egentlig ('really')	1

	und nun	1		
Adjectives	gut	2	fint	1
	sehr gut	1	utmerket ('excellent')	1
	schön	1	greit nok (lit. 'OK enough')	1
	sicher	1		
Adverb(ial)s	auf jeden Fall	1	iallfall	1
	jedenfalls	1	i hvert fall (hvertfall)	4
	bloß	1	i og for seg ('an und für sich')	1
	da	1	sannelig ('truly')	1
	dann	1	så ('then')	2
	trotzdem	1		
Interjections or exclamations	Ach	2		
	ach wirklich	1	du verden (lit. 'you world')	2
	aha	1	å	1
	großer Gott	1		
	hm	1		
Other	mag sein	1	hør her (lit. 'listen here')	1
	nicht direkt	1	jeg vet ikke ('I don't know')	1
	zugegeben	1	hun har jo rett ('she is MP right')	1
			tro? ('believe?')	1
Omission		36		30
Gap or other problem		3		4
Total		226		226

Table 3 shows that there is a wide range of correspondences in both languages. Similar categories of forms are used in Norwegian and German. Most often we find discourse particles. There are also: modal particles (note Norwegian *jo* and *da*); combinations of discourse particles and modal particles (again, *jo* and *da* in Norwegian); conjunctions, particularly the German and Norwegian equivalents of English *but*; adjectives that indicate acceptance; adverb(ial)s, particularly concessive expressions; interjections and exclamations, etc. There is further a substantial amount of zero correspondence or omission.

The findings are compatible with a similar study of *well* and its translations into Swedish and Dutch by Aijmer and Simon-Vandenbergen (2003). The affirmative response particle *ja* is a common translation of *well* in both of the Scandinavian languages (unlike German and Dutch). German is like Dutch in that most of the frequent translations have some relationship to words for 'now'; the main difference is the common use of *also*, which as far as I can see has no counterpart in the Dutch material. It is significant that similar means have been appropriated in different languages. We find both expressions of acceptance and concession, of agreement and disagreement, emotional expressions, etc. Some instances where the meaning is spelled out more explicitly are worth noting, in particular: the attention-getting forms *hör mal*,

nun sag mal, *hør her* (lit. 'listen here'); the hedging expressions *mag sein*, *nicht direkt*, and *jeg vet ikke* ('I don't know').

Aijmer and Simon-Vandenbergen suggest that *well* is "a heteroglossic option, accommodating the utterance to the context, in particular the hearer's expectations" (p. 1128). Whatever the correct interpretation may be, it is clearly the case that neither Norwegian nor German, nor the languages considered by Aijmer and Simon-Vandenbergen (Swedish and Dutch), possess fully-fledged counterparts of the English discourse particle *well*. Many different means are used to pick up facets of its meaning, and sometimes the meaning is lost altogether, or at least there is no explicit formal correspondence. Let's look at what further insight we can get by comparing the translations vs. the sources of *well*.

6 TRANSLATIONS VS. SOURCES

As the English-German-Norwegian corpus has not been built up to the same extent as the ENPC, the comparison here will be restricted to the ENPC and will focus on the correspondences of English *well* and Norwegian *vel*. These have the same origin, and both can be used as discourse particles.[4] Table 3 above shows that *vel* is used to translated *well* in about a fourth of the cases. But the two words are quite differently distributed in original vs. translated texts; see Figures 4 and 5. As the subcorpora of original and translated texts are equal in size (number of texts and approximate number of words), we can compare raw frequencies.

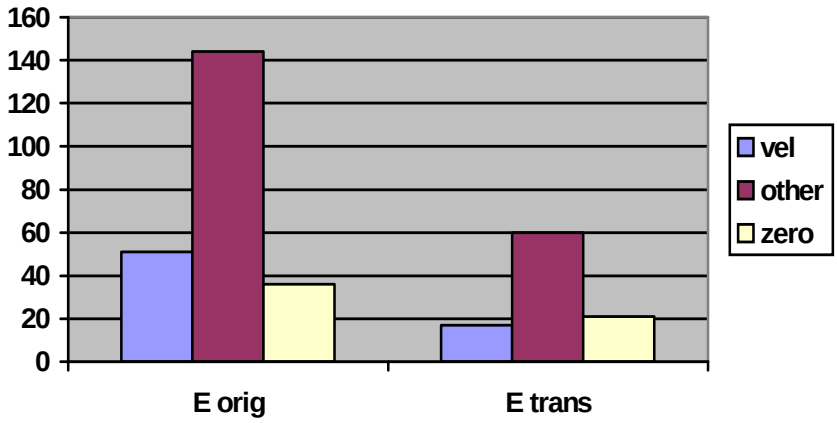

Figure 4 English well and its correspondences in the fiction texts of the ENPC

[4] Note that Norwegian *vel* can be both a modal particle and a discourse particle.

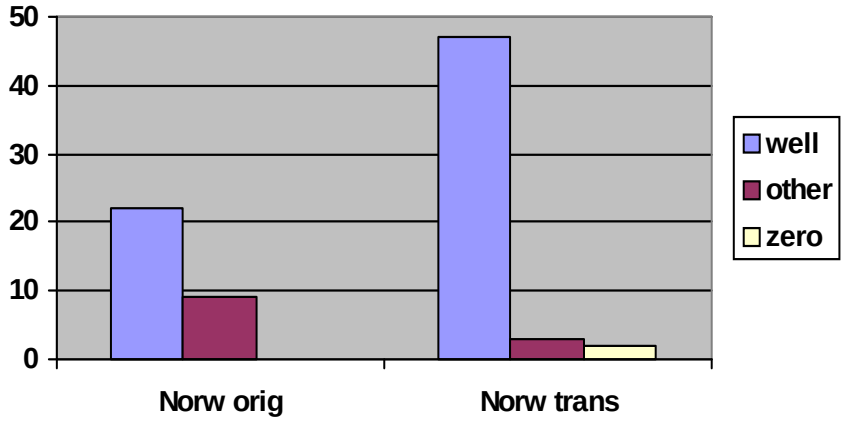

Figure 5 Norwegian vel and its correspondences in the fiction texts of the ENPC

A study of the correspondence patterns is instructive. *Well* normally corresponds to something else than *vel*, but *vel* corresponds closely to *well*, particularly in translation from English. It is not uncommon to find such an asymmetric cross-linguistic relationship. There is also a striking difference in the overall frequency patterns of the two discourse markers. *Vel* is far less common than *well*, with a total of 31 instances in Norwegian original texts and 52 in translations from English, as against 231 examples for *well* in English original texts and 98 in translations from Norwegian. In the case of *vel* the number goes up in the translations, due to the similarity to the English discourse particle *well*, which is frequent in the English source texts; for English *well*, we find the opposite relationship. Most clearly the difference is shown by the frequency of zero correspondence; for *well* this is common, in the case of *vel* there is hardly any zero correspondence.

To sum up, *well* and *vel* clearly overlap, but *well* has a wider range of use. This is why we get the distribution patterns shown in the figures. There is also a clear translation effect showing that translators are influenced by the linguistic choices in the source text. As there is no obvious counterpart, *well* goes down in frequency in translation from Norwegian.

7 PARALLEL TRANSLATIONS

To further explore *well* in a cross-linguistic perspective, let's look at some examples from our Norwegian-English-German-French corpus; see Figure 6.

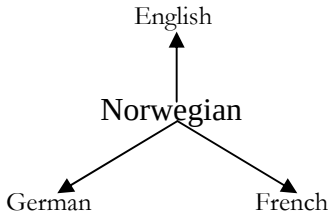

Figure 6 The Oslo Multilingual Corpus: Norwegian-English-German-French

The advantage of this sort of model, which we could call the star model, is that it makes it possible to compare translations across three languages, though there is no way of distinguishing clearly between language differences and translation effects. The most straightforward comparison here uses Norwegian as a starting-point, but I will continue with English well. I have searched for well in the English translations. The question is this: where the English translator has chosen well, what is the choice made by the German and French translators? These examples illustrate overt correspondences:

(4) Han finner det rådelig å vente litt med selve budskapet, og for å vinne tid, bøyer han seg ned og løsner i all stillhet snoren på skipssekken.
—*Jaså,* er det den karen som er ute og går, begynner skipperen i den vante gemenslige duren. (BHH1)

(4a) Er hält es für angebracht, mit seiner Mitteilung noch ein wenig zu warten, und um Zeit zu gewinnen, bückt er sich und bindet in aller Stille seinen Seesack auf.
"*Also*, hier sehen wir uns nun wieder", beginnt der Schiffer in seiner üblichen leutseligen Art.

(4b) He finds it advisable to put off his message a moment, and to gain time he bends down and quietly loosens the string of his duffel bag.
"*Well*, imagine you here," the skipper begins in his usual affable tone.

(4c) Il lui paraît raisonnable d'attendre un peu avant de délivrer le message lui -même et, pour gagner du temps, il se penche puis défait tranquillement les cordonnets de son sac.
"*Eh bien*, nous voilà donc de sortie, attaque le capitaine du ton familier qui lui est habituel.

(5) Jeg vet virkelig ikke.
— *Da* bare dropper vi spørsmålet nå. (JG3)

(5a) Ich weiß es wirklich nicht.
Dann lassen wir das für den Moment.

(5b) I really don't know.
Well, we'll let the question rest for the moment.

(5c) Je ne sais vraiment pas.
— *Bon*, laissons tomber cette question pour l' instant.

In (4b) *well* is used to start a conversation, in (5b) to round off a topic. We find that German and Norwegian tend to agree, while French is closer to English.

Sometimes *well* is inserted in the English translation, although both the Norwegian source text and the German and French translations lack a formal counterpart, as in:

(6) — Vi skal se på Franks herbarium, sa hun.
— Det synes jeg ikke du skal, repliserte Bill.
— Det synes jeg ikke du har noe med, sa Laura.

(6a) "Frank will mir sein Herbarium zeigen", sagte sie.
"Darauf solltest du verzichten", erwiderte Bill.
"Das geht dich nichts an", sagte Laura.

(6b) "We're going to look at Frank's herbarium," she said.
"*Well*, I don't think you should," Bill countered.
"*Well*, I don't think it's anything to do with you," Laura riposted. (JG3)

(9c) — Frank va me montrer son herbier, dit-elle.
— Je trouve que tu ne devrais pas, répliqua Bill.
— Je ne vois pas en quoi ça te regarde, rétorqua Laura.

(7) […] synes du at det er et altfor tilfeldig tema?
— Bare sett i gang, jeg går likevel ikke og legger meg før solen står opp. (JG3)

(7a) " […] hältst du das für ein zu schwammiges Thema?"
"Schieß los, ich geh sowieso erst schlafen, wenn die Sonne aufgeht."

(7b) " […] do you think it's too arbitrary a theme?"
"*Well* carry on, I shan't be going to bed before the sun comes up, anyway.

(7c) A moins que ce ne soit aussi un sujet trop superficiel pour toi...
— Vas-y, de toute façon je ne me coucherai pas avant le lever du soleil.

In (6b) the insertion of *well* makes the wording less categorical, taking into account the addressee's perspective. In (7b) the imperative is toned down. The need for such grease in the interaction apparently varies in different languages, as do the means of expression.

8 MULTIPLE TRANSLATIONS

Before I go on to the summing up, I will just give a further illustration from a corpus compiled according to the star model; see Figure 7.

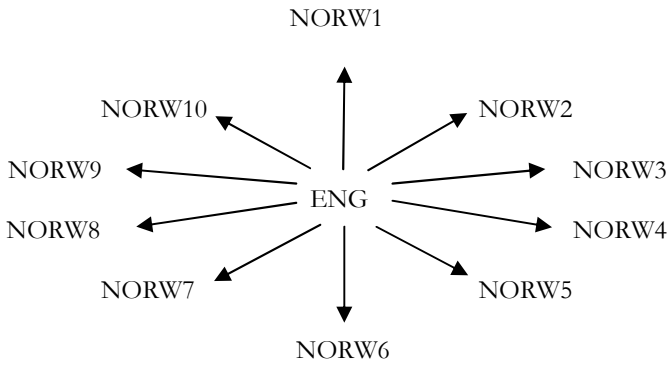

Figure 7 A multiple translation corpus: English-Norwegian

In this case we commissioned groups of professional translators to translate the same text independently. This is how a group of ten translators handled the first sentence in a short story by A. S. Byatt:

(8) IN THE MID-1980s, Bernard Lycett-Kean decided that Thatcher's Britain was uninhabitable, *a land of dog eat dog, lung-corroding ozone and floating money*, of which there was at once far too much and far too little. (Byatt, s. 1)

(8a) Midt i åttiårene fant Bernard Lycett-Kean ut at Thatchers England ikke var til å leve i, *det var for mange bikkjer om beinet der, fullt av lungetærende ozon og flytende pund* [lit. 'there were too many dogs about the bone there, full of lung-corroding ozone and floating pound'] som det både var for mye og for lite av. (transl. 1)

(8b) På midten av nittenåttitallet fant Bernard Lycett-Kean ut at Thatchers Storbritannia var ulevelig, *et land i jungellovens tegn, med ozon som tærte på lungene og flytende penger*, [lit. 'a country in the jungle-law's sign, with ozone which corroded on the lungs and floating money'] som det var både altfor mye og altfor lite av. (transl. 2)

(8c) Midtveis i 1980-åra slo Bernard Lycett-Kean fast at Thatchers England ikke var til å leve i, *et land av ulver, med lungetærende ozon og fri flyt av penger* [lit. 'a country of wolves, with lungcorroding ozone and free flow of money'] som det på en gang var altfor mye og altfor lite av. (transl. 3)

(8d) Mot midten av 1980-årene fant Bernard Lycett-Kean ut at Thatchers England ikke var til å leve i, *preget som det var av alles kamp mot alle, lungeetsende oson og flytende valuta* [lit. 'marked as it was by everybody's fight against all, lungcorroding ozone and floating currency'] som det på samme tid var enten altfor mye eller altfor lite av. (transl. 4)

(8e) Midt på åttitallet fant Bernard Lycett-Kean ut at Thatchers Storbritannia var ulevelig, *et nådeløst sted med lunge-etsende oson og raske penger*, [lit. 'a merciless place with lung-corroding ozone and quick money'] som det samtidig var altfor mye og altfor lite av. (transl. 5)

(8f) På midten av 1980-tallet var Bernard Lycett-Kean kommet til at Thatchers Storbritannia ikke lenger var til å bo i – *et land hvor livet var blitt det reneste bikkjeslagsmål, hvor ozonet åt opp lungene på folk, og pengene fløt i fritt fall*, [lit. 'a country where life had become the purest dogfight, where the ozone ate up the lungs of people, and the money floated in free fall'] altfor mye og altfor lite på en gang. (transl. 6)

(8g) På midten av 80-tallet bestemte Bernard Lycett-Ken seg for at Thatchers England var ubeboelig. *Det var et land der alle var ute etter hverandre, med luft som etset opp lungene og fri flyt av penger*, [lit. 'It was a country where all were out after one another, with air which corroded up the lungs and free flow of money'] som det forøvrig var både altfor mye og altfor lite av. (transl. 7)

(8h) En gang på midten av 80-tallet fant Bernard Lycett-Kean ut at Thatchers Storbritannia ikke var et levelig sted. *Det var et land av alle mot alle, av lungetærende ozon og penger i fri flyt*, [lit. 'It was a country of all against all, of lungcorroding ozone and money in free flow'] penger det både var altfor mye og altfor lite av. (transl. 8)

(8i) Midt på 80-tallet bestemte Bernard Lycett-Kean seg for at Thatchers Storbritannia var et ubeboelig land, *hvor den sterkestes rett rådet, ozon etset lungene, og det var fri flyt av penger*, [lit. '(un uninhabitable country,) where the right of the strongest prevailed, ozone corroded the lungs, and there was free flow of money'] som det samtidig var altfor mye og altfor lite av. (transl. 9)

(8j) Midt på 1980-tallet fant Bernard Lycett-Kean ut at Thatchers Storbritannia var blitt ubeboelig, *et barbarisk, kannibalistisk samfunn med dødelig oson og flytende penger*, [lit. 'a barbaric, cannibalistic society with deadly ozone and floating money'] som det både var altfor mye og altfor lite av på samme tid. (transl. 10)

Although all the translated versions differ greatly, the opening and the end of the sentences agree apart from minor differences in structure and word choice. The challenge is the noun phrase *a land of dog eat dog, lung-corroding ozone and floating money*. Two translators (7 and 8) started a new sentence, and most of them introduced one or more clauses. Such clause building is commonly found in translation from English into Norwegian (far more often than its opposite, i.e. clause reduction), in part due to structural and stylistic differences between the languages, in part probably also reflecting a tendency towards explicitation. The greatest problem is finding a way of handling the metaphorical *a land of dog eat dog*, which has no exact counterpart in Norwegian. Some translators found a metaphorical expression (transl. 1, 2, 3, and 6), others opted for descriptive phrases or clauses (transl. 4, 7, 8, 9), and two reduced the phrase to adjectives which capture part of the meaning of *dog eat dog* (transl. 5 and 10).

It is not easy to decide which is the most successful of the renderings – they have all been produced by experienced professional translators who had received prestigious prizes – but there is undoubtedly a lot of good material here for the study of translation and the training of translators.

9 CONCLUDING REMARKS

It is time to sum up. The types of multilingual corpora which I have taken up allow different ways of looking. Translation patterns give evidence both of language contrasts and of characteristics of translation. What we observe are correspondences. These correspondences have to be interpreted. The correspondences for the Norwegian modal particle *nok* and the English discourse particle *well* can be interpreted as evidence of their meaning. At the same time we find evidence of translation effects. These can best be identified with a corpus built according to the ENPC model. This is why we claim that it is appropriate both for contrastive analysis and for translation studies.

The systemic functional model has turned out to be a very useful tool in interpreting data from our corpora. Recently a student at the University of Oslo wrote an MA thesis on correspondences of the English verb *see* and its Norwegian cognate *se*, using the framework of process types from systemic functional linguistics. Two additional MA students have recently embarked on similar projects; one of them examines the Norwegian posture verb *stå* ('stand') and its correspondences in English and Italian; the other student looks at English *go* and its correspondences in Norwegian and German. My colleague Hilde Hasselgård has written a number of papers examining sentence openings in English and Norwegian in the light of the textual metafunction. A brand new Ph.D. thesis by Berit Løken (2007) explores expressions of possibility in English and Norwegian. The systemic functional model is used as a tool, but it is also enriched in the course of the cross-linguistic study.

The studies that have been carried out so far have convinced us that we can learn a great deal by examining multilingual corpora. But multilingual corpus studies are in their infancy. We need more multilingual corpora. We need different kinds of corpora. We need carefully constructed corpora. We need diamonds. We need stars. Above all, we must learn to use corpora in an insightful manner. We must learn to see more. And one of the ways is through corpora.

REFERENCES

Aijmer, Karin and Bengt Altenberg. 1996. Introduction. In Karin Aijmer, Bengt Altenberg, and Mats Johansson (eds), Languages in Contrast. Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994, 11-16. Lund Studies in English 88. Lund: Lund University Press.

Aijmer, Karin and Anne-Marie Simon-Vandenbergen. 2003. The discourse particle well and its equivalents in Swedish and Dutch. Linguistics 41: 1123-1161.

Hansen, Silvia. 2003. The Nature of Translated Text: An interdisciplinary methodology for the investigation of the specific properties of translations. Saarbrücken Dissertations in Computational Linguistics and Language Technology. Vol. 13. Saarbrücken: German Research Center for Artificial Intelligence and Saarland University.

Johansson, Stig. 1998. On the role of corpora in cross-linguistic research. In Stig Johansson and Signe Oksefjell (eds), Corpora and Cross-linguistic Research: Theory, method, and case studies, 3-24. Amsterdam & Atlanta, GA: Rodopi.

Johansson, Stig. 2006. How well can well be translated? On the English discourse particle well and its correspondences in Norwegian and German. In Karin Aijmer and Anne-Marie Simon-Vandenbergen (eds), Pragmatic Markers in Contrast, 115-137. Amsterdam: Elsevier,

Johansson, Stig. 2007. Seeing Through Multilingual Corpora: On the use of corpora in contrastive linguistics. Amsterdam & Philadelphia: Benjamins.

Løken, Berit. 2007. Beyond Modals: A corpus-based study of English and Norwegian expressions of possibility. Acta Humaniora No. 296. Faculty of Humanities, University of Oslo.

Neumann, Stella. 2003. Textsorten und Übersetzen. Eine Korpusanalyse englischer und deutscher Reiseführer. Frankfurt u.a.: Peter Lang Verlag.

Sinclair, John M. 1966. Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robins (eds), In Memory of J. R. Firth, 410-430. London: Longman.

Sinclair, John M. (editor-in-chief). 1987. The Collins COBUILD English Language Dictionary. London & Glasgow: Collins.

Sinclair, John M. and Malcolm Coulthard. 1975. London: Oxford University Press.

John M. Sinclair, John M., Susan Jones and Robert Daley. 2004. English Collocation Studies: The OSTI Report. Ed. by Ramesh Krishnamurty, including a new interview with John M. Sinclair, conducted by Wolfgang Teubert. London & New York: Continuum.

Steiner, Erich. 2005. Explicitation, its lexicogrammatical realization, and its determining (independent) variables – towards an empirical and corpus-based methodology. SPRIK report 36. http://www.hf.uio.no/forskningsprosjekter/sprik/docs/

Teich, E. 2003. Cross-linguistic Variation in System and Text. Berlin & New York: Mouton de Gruyter.