Pietilä, P. & O-P. Salo (eds.) 1999. Multiple Languages – Multiple Perspectives. AFinLA Yearbook 1999. Publications de l'Association Finlandaise de Linguistique Appliquée 57. pp. 105–134.

CORPUS STUDIES IN APPLIED LINGUISTICS

Kay Wikberg, University of Oslo
Stig Johansson, University of Oslo
Anna-Brita Stenström, University of Bergen
Tuija Virtanen, University of Växjö

Three samples of corpora and corpus-based research of great interest to applied linguistics are presented in this paper. The first is the Bergen Corpus of London Teenage Language, a project which has already resulted in a number of investigations of how young Londoners use their language. It has also given rise to a related Nordic project, UNO, and to the project EVA, which aims at developing material for assessing the English proficiency of pupils in the compulsory school system in Norway. The second corpus is the English-Norwegian Parallel Corpus (Oslo), which has provided data for both contrastive studies and the study of translationese. Altogether it consists of about 2.6 million words and now also includes translations of English texts into German, Dutch and Portuguese. The third corpus, the International Corpus of Learner English, is a collection of advanced EFL essays written by learners representing 15 different mother tongues. By comparing linguistic features in the various subcorpora it is possible to find out about non-nativeness generally and about problems shared by students representing different languages.

1 INTRODUCTION

1.1 Corpus studies and descriptive linguistics

Corpus-based language research has now been with us for more than 20 years. The number of recent books dealing with corpus studies (cf. Biber et al. 1998, Kennedy 1998, Granger (ed.) 1998a) and the recent publication of the *International Journal of Corpus Linguistics* are proof of the established status of this field of study. Two developments are particularly

worth noting in this context. One is the rapidly increasing size of corpora, the other is the development of new specialized corpora.

The most immediate benefit we have had of very large corpora so far is no doubt the new generation of monolingual English dictionaries that they have spawned. At the same time such corpora have provided us with a widened empirical basis for research into grammar, lexis, text types, and language variation. As far as grammar is concerned, it may be too early to judge to what extent access to megacorpora like The Bank of English and The British National Corpus (BNC; http://info.ox.ac.uk/bnc) will add information to what we know already, but it is obvious that one can expect them to contain more instances of what is rare in one-million-word corpora. Also, large corpora provide data which have not been available previously, such as spoken data classified in terms of domain (educational, business, institutional) and the regional distribution, age, and sex of the demographic respondents, as in the BNC. Similarly, large corpora are indispensable for the study of collocability, and should also allow us to say more about the distinctive properties of specific text types. Large corpora open up the possibility of doing statistical language analysis on an unprecedented scale. However, the snag with very large corpora is that their mere size may be deceptive. The output of any concordancing programme is always dependent on the composition of the corpus and the quality of the included texts.

The compilation of specialized corpora is due to the needs of special areas of research and the awareness of researchers that corpora will provide new types of data. It is corpora belonging to this category that we shall be concerned with in this paper. One such corpus which has taught us a great deal about the English spoken by young Londoners is COLT (The Bergen Corpus of London Teenage Language). This subcorpus of the BNC is basically a source for sociolinguistic research, but it has proved to contain interesting data on diachronic change in Present-day English as well. COLT has given rise to two more projects having to do with teenage language, i.e. UNO, ‘Språkkontakt och ungdomsspråk i Norden’, and EVA, ‘Evaluation of English in Schools’ (see Section 2).

1.2 Corpus studies and applied linguistics

It is natural that people involved in language teaching should have taken an interest in corpus-based research. It is not just that language teachers and textbook writers are potential consumers of corpus data, i.e. they need

authentic texts and fresh information on current usage. They also need information that may throw light on the second language learning process. Admittedly, such information has been available for a long time, but machine-readable texts with accompanying software enable us to do such research more systematically and exhaustively than before. Another reason why corpus studies are of interest to language teachers is the increased access to computer corpora, even in the classroom.

One basic type of applied linguistic research that had its heyday in the 60s and the early 70s is contrastive studies. Recently there has been a growing interest in transfer (cf. Odlin 1989) at the same time as contrastive studies have had something of a revival due to the development of parallel corpora and new techniques of aligning sentences on computer in two or more languages. This development has also attracted researchers in translation studies. The project presented in this paper (see Section 3) is the English-Norwegian Parallel Corpus, initiated by Stig Johansson. Actually, the name of the corpus is now slightly misleading, since the originally bilingual corpus has been expanded into a multilingual translation corpus. The corpus has so far been used as a source for a number of articles and contrastive M.A. theses in Norway and has engendered contrastive research in other Nordic countries as well.

The study of learner language got off the ground at the time when the interest of contrastive linguists started shifting from grammar towards pragmatics and text linguistics. Learner language research is constantly of great interest, but has long been in need of empirical data for the systematic study of cross-linguistic research on language learning strategies. The design of the International Corpus of Learner English (ICLE) is an attempt to provide such data (see Section 4).

2 CORPORA AND YOUTH LANGUAGE

(Anna-Brita Stenström)

Three projects are going on at the Department of English at Bergen University, all involving youth language and corpus-based research:

- COLT[1] (The Bergen Corpus of London Teenage Language)

[1] COLT has been sponsored by the Norwegian Research Council, the Norwegian Academy of Science, the Meltxer Foundation, and the Faculty of Arts at Bergen University.

- UNO[2] (Språkkontakt och Ungdomsspråk i Norden)
- EVA[3] (Evaluation of English in Schools)

2.1 Project descriptions

2.1.1 COLT (http://www.hd.uib.no/colt/)

The main purpose of the COLT project has been to collect a relatively large corpus of teenage language and make it available to students and researchers worldwide, the reason being that English teenage language was felt to be surprisingly little investigated. The corpus was collected in a short period in 1993. More specifically, 30 student volunteers, 13–17 year old boys and girls from socially different school districts in London, were provided with a Sony Walkman recorder and a lapel microphone and asked to record as many conversations as possible during a few days, preferably with their friends of the same age and in as many different situations as possible.

This resulted in a corpus of roughly half a million words, which is reasonably large, considering that the material is spoken. The tape-recordings were transcribed orthographically by professional British transcribers and later submitted to automatic word class tagging, carried out by a British research team. A large part of the material has also undergone a simple prosodic analysis, where utterances are chunked into tone units and marked for nucleus and intonation contour. The prosodic analysis was undertaken by student research assistants in Bergen.

In order to make the material easily accessible to all scholars interested in teenage talk, we first exposed the orthographic transcription with a search program on the Internet. Currently the COLT material is being produced on CD-ROM, a version where the orthographically transcribed, word-class tagged and prosodically analysed text files are accompanied by sound files and a user-friendly search program. The CDs will shortly be obtainable from the Norwegian Computer Centre for the Humanities at Bergen University.

---

[2] UNO is psonsored by the Nordic Council of Ministers.

[3] The EVA project has been comissioned by the Norwegian Ministry of Education and Research.

2.1.2 UNO (http://www.uib.no/uno/)

The UNO project, which is an offshoot of the COLT project, was launched in 1997. It is both a research project and a network, based on the collaboration of Nordic students and researchers. The purpose of the project is to investigate and compare the spoken (and informal written) language of Nordic teenagers in light of international, especially English, influence and inter-Nordic contacts. The focus is on the teenage vernacular in the five Nordic capitals, with some exceptions.

Corpora of teenage talk have been collected in all the Nordic countries except Iceland, which joined the project only recently. Since some of the corpora had already been compiled when the project started, they differ to some extent as regards year of collection as well as size and content. Therefore, supplementary recordings have later been made to make the corpora comparable. The most recent corpus, the Norwegian one, was recorded in Oslo in 1997 and 1998, mainly following the COLT model, but more systematically. Presently, we are hoping that an Icelandic corpus of teenage talk will follow.

The existing corpora have already been or are in the process of being transcribed orthographically. Eventually the materials from all five countries will be gathered in a common database administered at the Norwegian Computer Centre in Bergen, where it will be made accessible.

2.1.3 EVA (http://kh.hd.uib.no/eva/)

The EVA project was commissioned to the English Department at Bergen University by the Norwegian Ministry of Education in 1993. Its central purpose has been to develop material for teachers at different levels of the compulsory school system, to use in order to assess the language ability of their pupils. The material that has been developed responds to the following needs:

- a way of diagnostically assessing pupils' ability over a range of skills
- criteria for assessment, applicable at all levels of formality
- a way of documenting ability (especially oral)
- a way of enhancing oral skills through classroom practice

The first set of assessment material, intended for 14–15 year olds, was ready for use in 1996, and material for 11–12 year olds will be launched in 1999. The material is characterized by its profiling potential of a wide range

of pupils' abilities. Descriptors, profile forms, self-assessment forms and teacher observation forms provide a means of documenting pupils' ability in both test and non-test situations, hence of tracking progress round the year.

The development phases of each of the two sets of EVA material has involved extensive trialling in a cross section of Norwegian schools. The pupil material collected during the trialling section has been compiled into two electronic corpora. The first EVA corpus consists of both written and spoken secondary school language as well as a spoken language control corpus supplied by native speaker pupils. This corpus is available at http://129.177.24.28/eva/. The second EVA corpus, of written and spoken primary school language, should be ready to be used by the end of 1999.

2.2 Corpus-based research

All three projects have provided rich materials for research. Some of the research carried out so far on the basis of the corpora is summed up below.

2.2.1 COLT

The COLT conversations have yielded data for a large number of M.Phil. theses and a few Ph.D. dissertations at Bergen University. In addition, we know from numerous enquiries that the COLT material is being used for papers and more extensive studies by students elsewhere in the world.

One thing that immediately struck us, when we began studying the COLT conversations, was the frequent use of swearwords, which tended to abound when the speakers realized, the boys in particular, that the conversation was being recorded (e.g.*You're allowed, fuck off, you're allowed to say that. You're allowed to swear as much as you like*). This does not indicate, however, that the boys, in general, swore more than the girls. Factors such as age, social background and type of taboo word have more to do with it than gender. The number of swearwords, totally speaking, in the girls' talk was more or less equal to that of the boys, but they differed in type. The boys favoured strong swearwords (*fuck/ing*, *bloody* and *shit*), while the girls generally used weaker ones (*my god*, *goodness*; cf. Stenström 1995, Bynes 1998). The youngest boys (and girls)

rarely swore at all; instead, they uttered expressive ‘sounds’, such as *aargh, uuuhu, wooh*, etc.

Another thing that struck us was the frequent occurrence of pragmatic particles. Some were used as discourse markers, with a ‘bracketing’ function (Schiffrin 1987: 36–37), which help the speaker organize his/her turn (Stenström 1994: 63); others were used as interactional signals, which move the conversation forward, e.g. by appealing for or giving feedback (Stenström 1994: 61). The reduced form *cos* (from because) belongs to the first category. In the London teenage vernacular, *cos* is more often used as a discourse link, or ‘take-off’ for further talk, with no syntactic connection to the immediately preceding clause in terms of subordination, while the role of subordinating conjunction is more and more restricted to the form *because* (Stenström 1996, 1998). Another well-known item, frequently used as a discourse marker, is *you know*, which is not only used for that purpose, however, but also as an interactional signal, notably ‘empathizer’, by which the speaker ‘includes’ the listener in the ongoing talk (cf. Stenström 1994: 126–128). In the London teenage talk, a similar role is played by *yeah*:

> John: who's making a squeaking noise oh is this stupid thing. v>mimicking squeaking noise</nv>. Oh God don't ask her. ...(5) All it is yeah, is a project yeah that six peo= me and other five other people yeah in the school were asked to do yeah for a university which is studying ch= erm children's language, yeah and what it's like and basically I've got to carry it on me for a weekend yeah, record loads of different conversations on ten different tapes

This example shows very clearly that pragmatic particles generally do more than one thing at once; *yeah* serves both as a discourse maker (bracketing) and as an interactional signal (empathizing). Other items used for the same functions, by adults as well as teenagers, are OK and *right*. These items may or may not have an appealing effect and elicit oral feedback. An item that has attracted a great deal of attention in this connection is *innit* (from *isn’t it/ain’t it*; cf. Stenström & Andersen 1996, Berland 1997). This form is used by the teenagers as an ‘invariant tag’ (just like *OK, right* and *yeah*); that is, it is tagged on to a clause regardless of whether it agrees with the verb in the main clause or not. Thus we find, for instance, *You’re so dumb, innit* and *He gets upset quick innit*, besides the regular *It’s too bad innit*, where the tag agrees with the verb as well as the subject in the main clause.

Another item is *like*, which is not only extremely common among the London teenagers, especially middle-class girls, but also very versatile. Andersen (1998), who adopts the relevance-theoretic approach, describes it as a 'looseness marker', which points to some discrepancy between what the speaker actually says and what he has in mind (*like twenty minutes; if someone wanted to be like a doctor*), a function that others refer to as 'hedging', which reduces the force of an utterance (Stenström 1994: 128, Holmes 1995: 26). *Like* is also used by the London teenagers to mark reported speech (*he was like can I have breakfast*). We recognize at least some of the different uses of *like* met with in the London teenage talk in the Danish *lissom/ligesom* and the Norwegian and Swedish *liksom*, which, in the same way as like, have a definite tendency to be overused. Incidentally, in teenage talk, *like* seems to be on the point of outdoing both *sort of* and *kind of* in the hedging function (cf. Monstad 1998).

It has often been claimed that female speakers are less assertive than male speakers and therefore are more keen to use modal expressions as a safeguard. Mosaker (1998), who studied how the discourse marker *I think* is used by the teenagers in COLT and the adult speakers in the BNC (The British National Corpus[4]), found, contrary to expectations, that the male speakers in both corpora uttered *I think* more often than the female speakers. This does not necessarily mean, of course, that the male speakers were more tentative than the female speakers. The use of *I think* does not always reflect doubt and uncertainty, and male speakers may just use it differently.

Finally, the teenagers' use of intensifiers is worth mentioning, in particular their unexpected use of *well* as an adjective intensifier and *enough* as a premodifying instead of postmodifying intensifier. Examples such as *well cool, well hard, well nice* and, even more unexpectedly, *enough crap, enough gormless*, where both *well* and *enough* can be replaced by the prototypical adjective intensifier *very*, occurred in the corpus. Our first hunch was that this is an entirely new phenomenon, but it turned out that this usage was widespread as early as the 8th and 9th centuries. What is interesting, however, is that it fell out of use before the end of the 19th century (*The Oxford English Dictionary* 1989: 115 ff). The most likely reason why this usage is reappearing in today's London teenage vernacular is probably a matter of analogy and, in the case of *well*, related to its great versatility (cf. Stenström in press).

---

[4] The BNC consists of 10 million words of spoken English produced by adult speakers in the whole of Britain.

2.2.2 UNO

At the outset of the UNO project, we decided to focus on two areas of research: lexical analysis (the study of slang and informal loanwords) and discourse analysis (the study of discourse markers and speech styles). By and large, the original plan has been followed, although the field of research has gradually been widened to include, for instance, studies of attitudes to language and language identity.

A great deal of time and effort has been devoted to the slang study. A slang questionnaire was worked out by Ulla-Britt Kotsinas on the basis of a pilot study and containing 55 words, for which pupils aged 14 to 17 were asked to provide as many slang synonyms as possible. The questionnaire was distributed in all the Nordic countries except Iceland, which joined the project later. The words were selected on the basis of the following criteria:

- they triggered a large number of answers in the pilot study
- they triggered English synonyms in the pilot study
- they triggered numerous synonyms because they were ‘dirty’ words

All the data has now been collected and fed into a common database and is ready to be analysed in detail. A preliminary analysis (Kotsinas forthcoming) indicates that the influence from English teenage slang is considerable, but also that slang words from contact languages in general crop up in the local language. An earlier study of slang in Finland (Lainio 1997) showed that the source of slang words among Sweden-Finnish teenagers was both Swedish and English. Lainio emphasizes the important function of slang as a unitary factor, which is crucial for group cohesiveness. The impact of immigrant language is highlighted in Aasheim’s (1997) article on ‘Kebab-norsk’, where we learn that young people in the eastern part of Oslo, especially boys, tend to adopt slang words from immigrant groups, mostly due to the fact that most immigrants live in that area. Most of the slang words that Aasheim found are related to the police, drugs and sex.

As regards the second area of study, discourse analysis, Hilmisdottir & Wide (forthcoming) have studied the use of the Icelandic discourse marker *sko* in radio programmes for teenagers. They found that the description of *sko* in dictionaries, where it is described as an interjection and a marker of emphasis, does not correspond to its functions in real life, at least not judging by the teenage programmes on the radio. In their material, *sko* serves as a discourse marker affecting the structure of the discourse, and not as an interjection and marker of emphasis. Moreover, it is more frequently placed at the end of an utterance than at the beginning, as

the dictionaries claim. Another study within the area of discourse analysis, Guldbaek-Ahvo (forthcoming), shows that the use of quotation markers varies considerably in Danish teenage everyday conversations. She found that the students used both motion verbs and discourse markers to introduce a quotation, and often more than one at a time. Direct quotes, she noticed, were used to achieve a dramatizing and evaluating effect.

A third area that has been studied by network participants is language and identity and language and attitudes. Two studies of language and identity are based on data from 8th grade Danish-Turkish bilinguals in Koge school, Denmark. Quist (forthcoming) shows by means of a sociogram that boys and girls gathered in two separate groups and, moreover, that there was no ethnic dividing line in the boys' group, while the girls tended to keep to their own ethnic groups. It also shows that the social grouping was less fragmented among the boys, with few subgroups, while the girls gathered in a number of subgroups. The other study, reported in Møller (forthcoming) concentrates on the choice of language among three bilingual girls. Møller's hypothesis was that the shift between Danish and Turkish would be related to the girls' feeling of identity. What he found was that two of the girls distanced themselves from the parent generation and their Turkish background, while the third girl showed a keen interest in her Turkish background but also had nothing against Danish.

A related study (Gunnarsdotter forthcoming) is based on two Swedish dialects, more exactly on the influence of the Göteborg dialect on the local dialect in Alingsås, a small town ca 50 km from Göteborg. She found that dialect changes were related to which of the following three categories 'fika, fotboll och frikyrkor' ('coffee talk, football and free churches') the informants favoured. The language of students who spent their free time talking to friends over a cup of coffee was most affected by the Göteborg dialect, while the students who devoted their free time to sports preferred the local dialect. The language of the free church students was somewhere in between, close to the regional standard.

Among other studies can be mentioned a study of argumentative sequences in student groups' discussions about music (Wirdenäs forthcoming), which shows that argumentative sequences occurred twice as often among the students who followed a theoretical programme as among students who followed a practical programme. Fremer (forthcoming) has observed that generic *du* (meaning 'man') in today's Swedish is more common among young people than among adults. Finally, a study of the influence of English on Finland-Swedish interactants in role plays (Forskåhl forthcoming) indicates that the boys used the English

single-word terms as if they were part of the Swedish language by using Swedish inflections and Swedish articles. As regards somewhat longer expressions, the boys shifted to English, stepping into the fantasy world by 'being' a role figure. In other words, English is very much part of the game.

2.2.3 EVA

Three of the studies based on the EVA material deal with the correlation between teachers' grades and students' performance. Two of the studies focus on grammar and the third on vocabulary. Tjerandsen (1995) studied sentence complexity in the writings of 14–15 year olds and found that the complexity was generally higher in writings that received the best marks, but also that the fact that a text had a high sentence complexity was no guarantee that it should also be a given a very good mark. In other words, sentence complexity could not be used as a criterion for distinguishing between good and bad pupils. Rather, what distinguishes good from bad writing is not sentence complexity but error frequency. Strand (1998) investigated the level of grammatical accuracy in the oral production of the 14–15 year olds. More precisely, by comparing the criteria of the EVA oral tests with teachers' marks, she tried to discover what students at different levels of proficiency can be expected to master. Moreover, by studying ten morphological categories, she tried to find out to what extent the pupils' accuracy scores were reflected in the teachers' grades. Her main conclusion is that students' mastery of different grammatical categories can indeed be used as criteria for teachers to differentiate between learner levels, and she emphasizes the importance of developing 'an empirically verifiable approach towards the evaluation of pupils' skills' (1998: 60).

Urdal (1995), who studied the vocabulary in 14–15 year olds' written texts, starts out with the hypothesis that the evaluation methods used by teachers will have an effect on the grades given. She found that the grade correlated in a significant way with number of words per composition, lexical variation and lexical density. On the other hand, there was also a strong correlation between the number of errors and grades, regardless of the length of the composition. As regards the use of computers for a study of this kind, she underlines that, although computers can help us perform an objective analysis of a text's voca-bulary, they cannot help us decide whether the words are used appropriately.

Pedersen (1995) stresses the importance of vocabulary in language teaching. She refers to it as 'a key area in SLA' (1995: 74), which should have an effect not only on syllabus design but also on the evaluation of the learners' performance. Moreover, she says, the teaching of vocabulary should have a stronger communicative orientation. This is strongly supported by Hasselgren (1999). One of the main findings in her PhD dissertation is that 'Smallwords count.' (1999: 281). She found that "the use, or rather non-use, of smallwords was a striking revealer of non-nativeness in pupils' speech" (1999: 263). Nativelike fluency is hardly possible without them. This is an important finding which should be taken seriously by language teachers, all the more so since smallwords generally occupy a very modest position indeed in language teaching.

2.3 Concluding remark

By this rather compact summary of current research into teenagers' speech and writing, I hope to have demonstrated the advantages of using corpora. Without corpus data, large-scale research of this kind is simply not feasible.

3 CORPORA AND CONTRASTIVE STUDIES

(Stig Johansson)

Although the interest in multilingual corpora has developed particularly in the last few years, there is an early precedent for the use of computer corpora in contrastive studies, viz. the bilingual Serbo-Croatian – English corpus deve-loped within the Serbo-Croatian – English Contrastive Analysis Project. Accor-ding to the project plan, this corpus was to consist of original Serbo-Croatian texts and their translations into English, and original English texts and their translations into Serbo-Croatian (Filipovic 1969). For the latter purpose, the project chose half a million words from the Brown Corpus, together with a translation into Serbo-Croatian produced as part of the project.

The special advantage of such a corpus, apart from the obvious gain following from the possibility of using computational tools for automatic searc-
hes and analysis, is that it goes a long way towards solving the problem of equivalence that has bedevilled much contrastive work. How do we know

what to compare? What is expressed in one language by, for example, modal auxiliaries could be conveyed by quite different means in another language. The paired texts reveal the interlingual identifications made by the translators, and the use of parallel corpora containing such texts could be regarded as the systematic exploitation of the bilingual intuition of the translators whose work is represented in the corpora.

Much later than the Serbo-Croatian – English Contrastive Analysis Project, James (1980: 178) reaches the conclusion that translation equivalence is the best available *tertium comparationis* for contrastive analysis, i.e. the best common basis for generalizations on cross-linguistic similarities and differences. In his recent book on Contrastive Functional Analysis, Chesterman (1998: 60) stresses the role of corpus studies as 'a good source of hypotheses' on cross-linguistic differences and, above all, as a method of hypothesis testing.

Before we go on to consider the way we may set up corpora for cross-linguistic studies, there is a need to stress another special advantage of corpora in contrastive studies. While contrastive studies in the past were particularly concerned with a comparison of (parts of) language systems in the abstract, corpora now provide us with the tools for comparing languages in use.

3.1 Types of multilingual corpora

A distinction is often made between two main types of corpora for use in cross-linguistic research:

- corpora consisting of original texts in one language and their translations into one or more languages – let us call these *translation corpora*;

- corpora consisting of original texts in two or more languages, match ed by criteria such as the time of composition, text category, in tended audience, etc. – let us call these *comparable corpora*.

It has been pointed out (e.g. by Lauridsen 1996) that we must use comparable corpora for cross-linguistic comparison. There are indeed a number of problems with translation corpora:

- the range of translated texts is restricted as compared with the range of original texts;

- there may be special features in translated texts, either due to source language influence (cf. Gellerstam 1996) or reflecting general features characteristic of translated texts (cf. Baker 1993) – the language of translation has even been characterized as a third code (cf. Øverås 1996);

- there may of course also be outright mistakes in translation.

In contrast, comparable corpora represent natural language use in each of the languages and they therefore provide a valid basis of comparison. But there are also problems with comparable corpora:

- How do we match texts across languages? If, for example, we want to compare informational density in German vs. Norwegian non-fictional prose (cf. Fabricius-Hansen and Solfjeld 1994), what criteria do we use for the selection of texts, to make sure that we compare equal entities?

- How do we know what linguistic features to compare? If, for example, we want to study how the notion of possibility is expressed in English and Norwegian (cf. Løken 1997), how do we identify the relevant forms?

In contrast, there is a perfect match of texts in translation corpora. Moreover, the forms in the original and the translation are intended to express the same meaning, and we can study the cross-linguistic correspondences established by skilled bilinguals.

Both types of corpora clearly have their advantages and disadvantages. Fortunately, it is not necessary to choose between them. Both can be combined within the same overall framework, as has been done in the English-Norwegian Parallel Corpus (ENPC). Each can then be used as a means of controlling and supplementing the other.

3.2 Structure and uses of the English-Norwegian Parallel Corpus

The structure of the ENPC is very much reminiscent of the corpus designed for the Serbo-Croatian – English Contrastive Analysis Project (see further Ebeling, forthcoming). There are original texts from both languages, together with their translations into the other language. The main difference is that both the originals and the translations included in the ENPC have been published, which should provide some standard of quality, as the translations have presumably gone through an editing process. Other differences (apart from the obvious difference in languages) have to do

with the types of texts included, the balancing of the texts across languages, the time of publication, and the total size of the corpus. All in all, the ENPC contains 200 texts, or about 2.6 million words in all. For a detailed description, see Johansson et al. (1999) or the web page of the project: http://www.hf.uio. no/iba/prosjekt/.

The structure of the ENPC is summarized in Figure 1. The boxes indicate the main components of the corpus, and the lines show the types of studies made possible by this type of design:

- contrastive studies based on parallel original texts (see one of the diagonal lines in Figure 1);
- contrastive studies based on original texts and their translations, going from source text to translation and/or from translation to sour ce text (see the horizontal lines in Figure 1);
- various types of translation studies, e.g. focusing on (a) translation problems viewed from either language (see the solid horizontal lines in Figure 1), (b) deviations of translated texts as compared with original texts in the same language (see the vertical lines in Figure 1), and (c) general features of translated texts (see one of the lines in Figure 1).

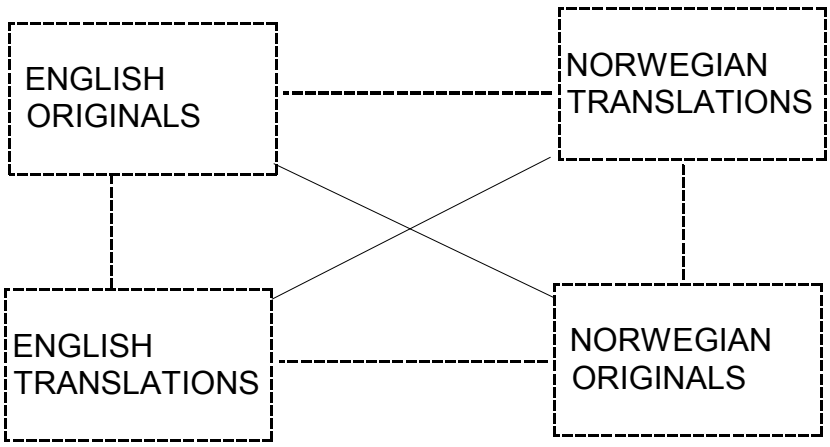

FIGURE 1. The structure of the English-Norwegian Parallel Corpus.

The corpus has already been used for a number of articles and MA theses on topics such as: modal auxiliaries, connectors, word order and thematic structure, and lexis (see the list of publications on our web page; cf. the web address given above). The most substantial study so far is the forthcoming doctoral thesis by Jarle Ebeling on presentative constructions in English and Norwegian.

An illustration of how the structure of the ENPC can be exploited both for contrastive analysis and translation studies is a recent paper on “Loving and hating in English and Norwegian” (Johansson 1998b). The starting-point for this paper was the observation of apparent over-use of Norwegian *hate* in translations from English. Figure 2 summarizes the results of a comparison of the frequency of *love* and *hate* and their Norwegian counterparts *elske* and *hate* in the fiction texts of the ENPC (30 original fiction texts from each language, with their translations into the other language). As the texts in each part of the corpus are balanced, we can compare raw frequencies.

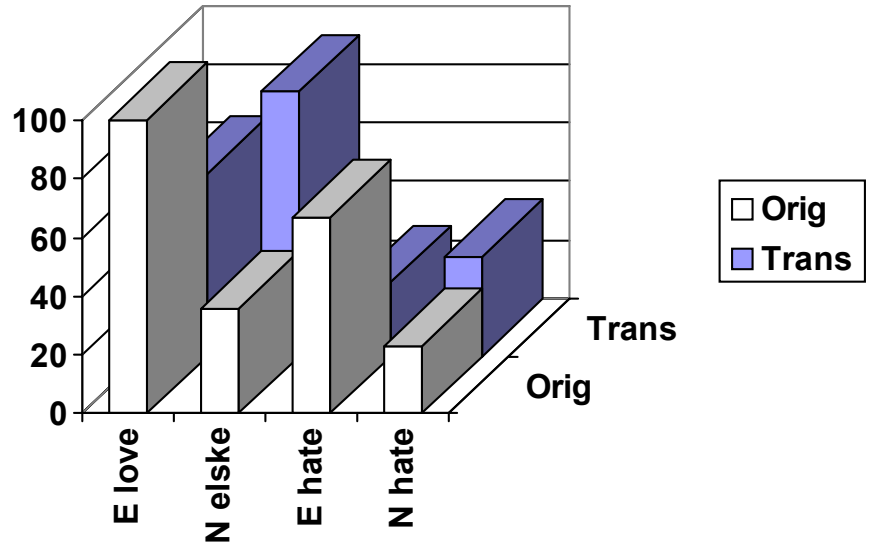

FIGURE 2. The distribution of English *love* and *hate*, and Norwegian *elske* and *hate* in original and translated fiction texts of the ENPC (30 texts of each type).

We see that there are sharp differences in the frequency of the English verbs and their Norwegian counterparts in original texts. The English verbs are about three times as common, testifying to a wider area of use. In the Norwegian translations we find a range of correspondences, far richer than in bilingual dictionaries. Nevertheless, the translators do not seem to be sufficiently aware of the differences between the languages. There is a

tendency to equate the English and the Norwegian verbs, as shown by a comparison of the frequency distributions for the verbs in translated text. See further Johansson (1998b).

3.3 Towards a multilingual corpus

If we want to gain insight into language and translation generally, and at the same time highlight the characteristics of each language, it is desirable to extend the comparison beyond language pairs. The ENPC project has therefore been extended to include translations of many of the English original texts into three other languages: German, Dutch, and Portuguese. Together with the Norwegian translations and the translations into Swedish and Finnish of related projects in Sweden (Lund/Göteborg) and Finland (Jyväskylä/Savonlinna), we can then compare across six languages using the English original texts as a starting-point (see Johansson 1998a).

The expansion has focused particularly on the triplet English-Norwegian-German, for which we are collecting translations into the other two languages for source texts in Norwegian and German as well as for English (a joint project with the Department of Germanic Studies and the Section for Applied Linguistics at the University of Oslo). The next section reports on a small-scale study based on a subcorpus of 16 English fiction texts and their translations into German and Norwegian.

3.4 The English verb *spend* and its correspondences in Norwegian and German

Among his examples of translationese in Swedish texts translated from English, Gellerstam (1996: 59) mentions the overuse of *tillbringa* corresponding to English *spend*. Gellerstam's study was, however, limited to a comparison of original and translated Swedish, and he had no access to the English original texts. In our case, we start from English and go to the German and Norwegian translations.

Both German and Norwegian have verbs similar to Swedish *tillbringa*, and they are frequently used as translations of *spend* plus a temporal NP. Other transitive verbs which sometimes occur in the translations are German and Norwegian verbs meaning ‘use’. Examples:

(1) He liked Sir Bernard Hemmings, but it was an open secret inside "Five" that the old man was ill and *spending less and less time in the office*. (Frederick Forsyth)
Er mochte Sir Bernard Hemmings, aber es war in "Fünf" ein offenes Geheimnis, daß der alte Mann krank war und *immer weniger Zeit im Büro verbrachte*.
Han likte Sir Bernhard Hemmings, men det var en åpen hemmelighet i "Fem" at den gamle mann var syk og *tilbrakte mindre og mindre tid på kontoret*.

(2) "Look Brian, I've *spent two years on that investigation*. (Frederick Forsyth)
"Hören Sie, Brian, ich habe *zwei Jahre auf diese Nachforschungen verwendet*.
"Hør nå, Brian. Jeg har *brukt to år på denne etterforskingen*.

More interestingly perhaps, we often find a restructuring of the clause, with the use of an intransitive verb plus an adverbial, as in:

(3) "But I *spent the night at Rose's*." (Jane Smiley)
"Aber ich hab *heut nacht bei Rose geschlafen*."
"Men jeg har jo *ligget over hos Rose*." (lit. 'lie over')

(4) Since the age of eighteen, he'd *spent an accumulated nine years in jail*. (Sue Grafton)
Seit seinem achtzehnten Lebensjahr hatte er *alles in allem neun Jahre im Gefängnis verbracht*.
Siden attenårsalderen hadde han *sittet inne i tilsammen ni år*. (lit. 'sit inside')

In (3) we note that the Norwegian translator has chosen a phrasal verb that lexicalizes the notion 'stay the night', while the German translator has an intransitive verb plus an adverbial. In (4) the German translator has opted for *verbringen*, and the Norwegian translator has picked a phrasal verb that lexicalizes the notion 'be in jail'.

A particularly interesting type of restructuring is found with the pattern *spend* + temporal NP + V-*ing*, a common pattern in the English original texts (close to half of the examples in the material). Examples:

(5) After leaving school at sixteen, Rawlings had *spent ten years working* with and under his Uncle Albert in the latter's hardware shop. (Frederick Forsyth)
Nach seinem Schulabgang im Alter von sechzehn hatte Rawlings *zehn Jahre* in der Eisenwarenhandlung seines Onkels Albert *gearbeitet*.
Rawlings hadde sluttet på skolen da han var seksten år og siden *arbeidet i ti år* sammen med og under sin onkel Albert som drev jernvarehandel.

(6) We *spent a lot of the time driving*, in our low-slung, boat-sized … (Margaret Atwood)

Die meiste Zeit fuhren wir in unserem niedrigen, bootsförmigen Studebaker herum …
Mye av tiden kjørte vi bil, en lav Studebaker, …

(7) He *spent pleasurable hours dithering* over questions of punctuation. (Anne Tyler)
Er *grübelte vergnügliche Stunden lang* über Interpunktionsprobleme nach.
Han *tilbrakte koselige timer med å gruble* over tegnsettingen.

Here again the translators often opt for an intransitive verb plus an adverbial, with the *ing*-verb 'raised' to the main clause (and without a verb corresponding to *spend*). But we also find translations with *verbringen* and *tilbringe*, e.g. in the Norwegian translation of (7), which has a structure which is more in agreement with the English original (but with an infinitive corresponding to the English *ing*-form).

The corpus gives a far richer picture of correspondences than bilingual dictionaries, though it is not possible to show this in detail here. For the language learner and the student of translation, there is a great deal to learn from a study of the correspondences in the corpus. But to what extent do the choices in the translations agree with the patterns we find in original texts?

A comparison of original and translated Norwegian, similar to Gellerstam's Swedish study, shows that Norwegian *tilbringe* is more than twice as common in texts translated from English than in Norwegian original texts. As there were not enough texts available, it was not possible to make the same comparison for German.

If we go now to Figure 3, we see examples of both overuse and underuse in translations. Apart from *tilbringe* in original and translated Norwegian, the figure compares the distribution of *keep* V-*ing* (as in *keep going*; cf. Davidsen Bjerga 1998) and *spend* in English original texts and in translations from Norwegian. While *tilbringe* is overused in translations from English, both *spend* and *keep* V-*ing* are less frequent in texts translated from Norwegian than in English original texts, presumably because of the lack of close formal correspondences in Norwegian. In all these cases, we see how the language of the translations has been influenced by the source language texts.

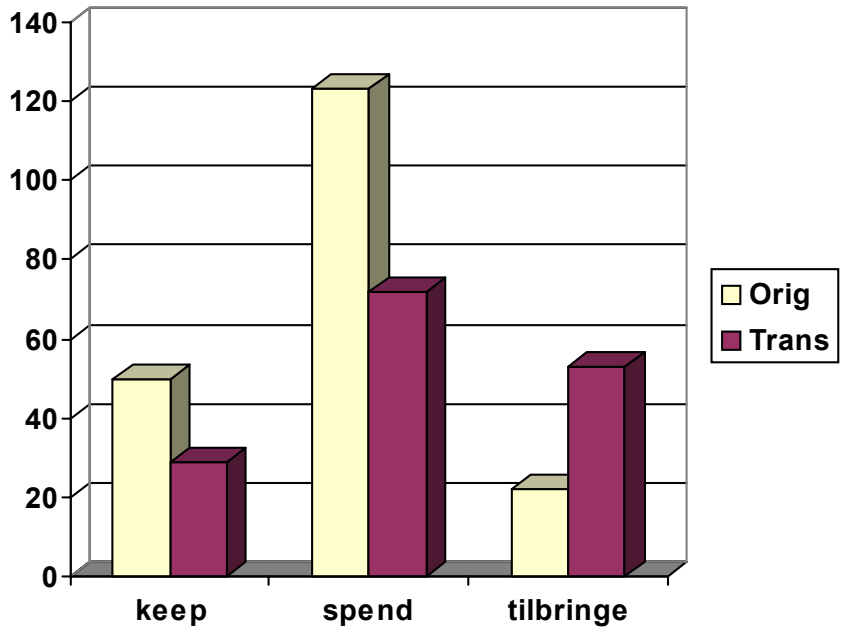

FIGURE 3. The distribution of English *keep* V-*ing*, English *spend*, and Norwegian *tilbringe* ('spend') in original and translated fiction texts of the ENPC (30 texts of each type).

3.5 Corpora and contrastive studies: Problems and prospects

As the examples above will have shown, translation corpora must be used with caution, but with the control functions built into the design of the ENPC we can get the advantages both of translation corpora and of corpora of original texts. The role of multilingual corpora must not be exaggerated, however. A corpus like the ENPC is for example too small for lexical studies beyond the core vocabulary, and it does not adequately represent the range of text type variation in the languages. Depending upon the topic, it will often be necessary to go beyond the corpus. This is as true of contrastive studies as of language research in general.

Corpora like the ENPC can be useful in teaching as well as in research. One of the biggest problems with grammars and dictionaries is going from description to language use. With the possibilities offered by electronic media, we can now imagine a grammar and a dictionary linked to a corpus. The dictionary and the grammar provide the description, the corpus the link to language use. Developing this kind of integrated teaching

tool will be one of the most challenging tasks for the future in applied contrastive studies.

4 WHAT CAN WE LEARN FROM CORPORA OF LEARNER LANGUAGE?

(Tuija Virtanen)

4.1 Introduction

A number of computer corpora consisting of EFL data are at present available or under compilation. The majority of these consist of written language, for obvious reasons. Yet, there are also projects which involve collection of data for corpora of learner speech. Corpora of learner language can be used by teachers and learners alike; they can form a basis for development of pedagogical tools and teaching materials, and they can be used in research on SLA.

My focus is on the International Corpus of Learner English (ICLE), which has relevance to members of AFinLA because of its present and future role in applied linguistics and because of the Nordic subcorpora included in it. After a brief presentation of ICLE, I will refer to some of the results originating in research in progress in which ICLE is being used and discuss some of the implications of these and other results. Finally, I will touch upon pedagogical applications of corpora of learner language. For information concerning the ICLE corpus, two very good sources are Granger (1998a) and the ICLE homepage at http://juppiter.fltr.ucl.ac.be/ FLTR/ GERM/ETAN/CECL/cecl.html .

4.2 ICLE – the International Corpus of Learner Language

Sylviane Granger from the University of Louvain-la-Neuve, Belgium, has launched a series of interesting corpus projects in applied linguistics. The one we are concerned with here, ICLE, is a corpus of advanced EFL writing, which will be available around the year 2000. There is also a small native-speaker corpus of similar student writing, which can be used for comparative purposes.

ICLE consists today of 15 different subcorpora, which are made up of advanced EFL essays of some 500–1,000 words each, written by

university students of English representing different mother tongue backgrounds. Each subcorpus consists of the minimum of 200,000 words; when completed, ICLE will thus consist of some three million words. One of the advantages of ICLE is the fact that it allows research into the written production of advanced learners of English, whose EFL proficiency is not as well documented as that of school-children at various levels. For information concerning the tagging of the corpus and other aspects of corpus compilation, see Granger (1998a); Virtanen (forthcoming).

Although small in size, ICLE is useful for a number of purposes because it is specialized according to six shared and seven variable features (see Granger 1998b: 9–10). To start with the shared features, the students are of approximately the same age, though in some countries such as Sweden university students are generally much older than elsewhere. The learning context is an English department and the level is advanced, though what constitutes an advanced EFL learner, of course, varies from one educational system to another. In Sweden, for instance, English is studied at university level during four 20-week terms. The medium is writing and the genre is student essay of a non-technical kind. Secondly, the variable features concern sex, mother tongue, region, command of other foreign languages, language practice, essay topic and task setting. The topics of the essays vary. They have been designed to invite argumentation; yet it is up to individual students to decide whether or not they will conform to the argumentative type of text. The task setting varies as to whether it is timed or untimed, examination, and whether or not the students have used reference tools such as dictionaries or grammars. This kind of background information will be coded in the corpus to allow automatic retrieval of, for instance, all essays written on a particular topic without reference tools in a timed examination setting by female students whose mother tongue is Finland-Swedish, who also know Finnish and French, and who have spent a year in an English-speaking country.

Three Nordic countries are represented in the project: Finland, with a completed subcorpus consisting of essays written by Finnish-speaking and Swedish-speaking Finns; Sweden, with a subcorpus close to completion; and Norway, just starting to collect essays for ICLE. The Finnish subcorpus

consists of essays originating in a number of different universities.[5] The Swedish subcorpus comprises essays written by students at the universities of Lund, Göteborg, and recently Växjö, and the national coordinators are Bengt Altenbe rg and Karin Aijmer. The Norwegian coordinator is Nancy Lea Eik-Nes (NTNU in Trondheim).

4.3 ICLE-based research in progress

What can we learn from corpora of learner language? The pilot studies reported in Granger (1998a) focus on a wide range of linguistic phenomena that can be investigated using computerized corpora of learner language. Most corpus studies concern learner lexis and grammar. However, even students of di scour se, pragmatics and rhetoric can, to an extent, benefit from such tools. 'Investigating discourse-pragmatic aspects of learner language on computer' will therefore be the theme of an AILA'99 symposium, organized by Tomoko Kaneko and Tuija Virtanen. The obvious outcome of corpus studies of learner language is a revival of contrastive linguistics in a new form, as shown, for instance, by the popularity of a recent conference in Louvain-la-Neuve on the theme 'Contrastive linguistics and translation studies: Empirical approaches' (5–6 February, 1999). Because of the large number of subcorpora, some of which represent non-Indo-European mother tongue backgrounds, ICLE is also particularly well suited for investigations of aspects of nonnativeness in general.

The first ICLE-based studies indicate for instance that advanced students of English show a clear preference for core vocabulary in their writing. This characteristic seems to be shared by students representing different mother tongues. Future reseach will show whether there are other aspects that are typical of non-nativeness, such as some of the avoidance tendencies manifested in these data. There are of course also characteristics which can be explained with reference to the students'

---

[5] The Finnish coordinators are Håkan Ringbom (Åbo Akademi University) and Tuija Virtanen (Växjö University). Signe-Anita Lindgrén (Åbo Akademi University) has been employed as the project assistant of the Finnish team. We are grateful to the following collaborators at Finnish and American universities for providing essays for the two corpora: Riikka Alanen, Andrew Chesterman, Ulla Connor, Roy Goldblatt, Helena Halmari, Mary Hatakka, Frank Hubbard, Pekka Hirvonen, Sirpa Leppänen, Carolyn Matalene, Anna Mauranen, Bo Pettersson, Outi Pickering, Anne Pitkänen-Huhta, Cathy Rohlich, Kari Sajavaara, Rachel Stewart, Kauko Timlin, Gregory Watson. Other people in these departments are also likely to have been involved in the collection of ICLE/LOCNESS data and deserve due thanks for their contribution.

mother tongue background. Non-native characteristics are generally discussed in terms of 'overuse' vs. 'underuse' in relation to native-speaker data. According to Ringbom (1998: 51), "it seems that the non-native features of the ICLE essays are less due to errors than to an insufficient and imprecise, though not necessarily erroneous, use of the resources available in English."

We can also distinguish a tendency for North European students to adopt characteristics of fairly informal spoken language and use these even in relatively formal settings such as essay writing. Altenberg (1998a) argues that this points to insufficient register awareness and stylistic competence in a foreign language. It also points to the heavy impact of spoken English in this part of the world. Petch-Tyson (1998) thus shows that the Nordic subcorpora manifest a high degree of writer/reader visibility. This tendency is confirmed by the findings of Ringbom (1998) as concerns core vocabulary and those of Virtanen (1998) as regards the use of direct questions in these data. Petch-Tyson's study shows that American students appear more detached from their texts and instead focus on message content. Preliminary results of my work on the progressive in ICLE suggest that American students express involvement with the text and the audience for instance through the use of this grammatical construction, which allows them to include a high degree of affect while focusing on the content (Virtanen 1998b).

In a recent paper on the use of the verb *make*, Altenberg (1998b) suggests that the Swedish students experience English to be fairly similar to their mother tongue. This would seem to be in contrast with the perception of English by the French-speaking students whose essays are included in ICLE (see also Granger & Altenberg, forthcoming).

Finally, Virtanen & Lindgrén (1998) study attitudes expressed by Swedish-speaking EFL students in Finland and Sweden concerning the use of the major geographical varieties of English in writing and match the results with comparable ICLE data to find out to what extent students do what they say they are doing in their written production.

Many of the studies referred to above start out from a lexical item or aim at grammatical or discourse-pragmatic phenomena through lexical cues simply because this is what the software available to linguists today allows us to do. This may result in a sort of backwash effect if carried out too far: We may choose to focus on what can be studied with the help of computer tools, rather than what we wish to study and what we find important to study. Also we may choose to quantify where possible and hence potentially give undue prominence to features that are readily

subjected to counts. Keeping these words of warning in mind, however, we can and must go on trying to develop better ways of studying corpus data for the purposes we want to, rather than those dictated by the computer. The advantages of corpus data are obvious: a large amount of data – or a smaller amount but relatively specialized data; the rapidity of the search procedure, allowing for an infinite number of repeated or modified searches; the fact that computers do not get tired and do the routine work again and again. The responsibility of designing the appropriate search procedure is ours and it will have to be very clearly documented in the methods and materials section of our reports of the results. Obviously, even at best the results are always only as good as the corpus itself.

Within the area of learner discourse, connectors are a popular object of study (see e.g. Granger and Tyson 1996; Altenberg and Tapper 1998). They can be relatively easily retrieved automatically and their use or misuse is likely to point, for instance, to influence of the learners' mother tongue or their lack of register awareness. Focusing on connectors and other explicit signals of cohesion can, however, only be a first step towards understanding what makes or fails to make texts coherent in a given context. Similarly, it is not enough to study lexical items present in the corpus to understand learner vocabulary. Ringbom (1998: 49) points out that

> even though transfer-based grammatical or lexical errors are not especially common in ICLE, this does not mean that the L1 plays an insignificant part in the form the texts have taken. It rather means that transfer is manifested in other, more subtle ways than the obvious errors made by learners at earlier stages. Most clearly this can be seen in learner groups whose L1 is structurally different from English. Evidence of "covert transfer" – i.e. "where L1-based procedures are being used in the absence of appropriate L2-procedures being available" and where "cross-linguistic formal similarity is thus largely irrelevant" – is manifested by avoidance or underuse rather than overuse: L2-constructions without direct equivalence in the L1 tend to be avoided or underused.

Avoidance is difficult to detect – but not necessarily more difficult than the occurrence of zero elements in a standard corpus (see e.g. Biber 1988). In other words, you can, for instance, get at deletion of the relative pronoun or *that* introducing a nominal clause by figuring out a search procedure which adequately describes the immediate context of the potential zero element. The task is obviously easier if the corpus is tagged. This kind of 'detective work' is a reality in today's corpus studies because of the slow development of the software.

To conclude, ICLE can to an extent be used to compare written production by advanced EFL learners representing different mother tongue backgrounds and it can definitely be used to study nonnativeness in general because of the wide range of mother tongue backgrounds. It is best suited for comparisons of learner lexis and grammar though local-level discourse phenomena can also be detected through lexical cues. As a spin-off effect, ICLE is also likely to reveal patterns that originate in different educational cultures. If completed by data from less advanced EFL learners and comparable native-speaker non-professional and professional writing in English and in the EFL students' mother tongue, it can be an important link for detecting longitudinal and cross-linguistic learner profiles. Finally, ICLE can serve as a generator of ideas and a testing ground for hunches, because of the easy access to comparable data from the different subcorpora.

Ideally, what we get out of corpus studies should help us detect emerging patterns not covered by existing linguistic models and theories. This, to me, would be the starting point of corpus linguistics, i.e. a new kind of linguistics based on emerging patterns in the data which have gone unnoticed in earlier approaches. In reality, however, we often focus on what we already 'know' or expect to be there, for instance through familiarity with the language skills of our students. Or we focus on what we can relatively easily retrieve from the corpus. We get immediate results and through quantification contribute to building up a myth of objectivity around those results. Even when critical of the results, we may, however, be satisfied by simply classifying the results in a number of categories, depending on the degree of delicacy we wish to adopt, and leave the description of the linguistic phenomenon under attention at that. At this point, it is clearly an advantage if we have a linguistic model or theory to link the results to, to be able to account for them in a systematic manner. Ideally, again, there is a two-way relation between theory and application; in practice, however, models and theories are adopted by applied linguists but the results of these studies do not necessarily contribute to the development of theoretical linguistics. There is, however, no reason why the results of corpus studies should not do just that: form a basis of new thinking in theoretical linguistics, too.

4.4 Pedagogical applications

Finally, access to learner corpora also contributes to the development of teaching materials and methods. Corpora of learner language allow teachers and learners to profit from authentic learner data to make the learning process more effective. Examples of such uses of learner language on computer also appear in Granger (1998a).

Language teachers have a solid experience and feel for what needs to be emphasized in EFL teaching at various levels. Yet, a corpus of learner language can reveal patterns which are unexpected to the language teacher and it can show the relative importance of the issues that teaching materials and teachers consider central in the syllabus.

Furthermore, corpus data can profitably be used as materials within the framework of problem-based teaching and learning, even if this may involve some editing of the data. Sensitizing students to patterns visible in KWIC concordances results in a more effective approach to learning. These data can be used for meaningful work in a study group where students solve a given problem and subsequently present their solution to their peers. It is useful for students to see the richness of analyses you can arrive at using the same corpus data. The procedure which demands that students form linguistic arguments to support the analysis they have arrived at using corpus data and to present and defend their analysis in the study group or classroom setting can be a basis for creating a more effective learning environment in higher education.

However, results of studies such as the ones reported above and the use of a corpus of learner language as a pedagogical tool immediately raise the tricky issue of a norm, which repeatedly demands attention in studies of learner language. We all know that there is no general consensus about what exactly constitutes native-speaker English. Instead of making implicit or explicit comparisons of learner language with a myth of an ideal Native Speaker, people using ICLE can compare EFL data, in general or by subcorpora, with similar materials written by native-speaker students in Britain and the U.S.A. In this way, they can compare non-professional EFL writing with non-professional native-speaker writing. When a comparison with professional native-speaker writing is desirable, it is possible for instance to use English-language editorials representing different kinds of newspapers and different parts of the English-speaking world. Exposing teachers and advanced EFL students to variation in native-speaker English is a good ground for insights and contributes to an effective approach to learning.

References

Aasheim, S. 1997. 'Kebab-norsk' – Fremmedspråklig påvirkning på ungdomsspråket i Oslo.In U-B. Kotsinas, A-B. Stenström & A-M. Karlsson (eds.) *Ungdomsspråk i Norden.* Stockholm: Meddelanden från Institutionen för nordiska språk vid Stockholms universitet MINS 43, 235–242.

Altenberg, B. 1998a. Exploring the Swedish component of the International Corpus of Learner English. In B. Lewandowska-Tomaszcyk & P.J. Melia (eds.) *Proceedings of PALC'97: Practical Applications in Language Corpora (Lódz, 12-14 April 1997).* Lódz: Lódz University Press, 119–132.

Altenberg, B. 1998b. Advanced Swedish learners' use of causative *make*: A contrastive background study. Paper presented at the International Symposium of Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, The Chinese University of Hong Kong, 14–16 December 1998. Extended abstract included in the symposium proceedings, 7–9.

Altenberg, B. & M. Tapper 1998. The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (ed.) *Learner English on computer*. London & New York: Addison Wesley Longman, 80–93.

Andersen, G. 1998. The pragmatic marker *like* from a relevance-theoretic perspective. In A. Jucker & Y. Ziv (eds.) 1998, 147–170.

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, and E. Tognini-Bonelli (eds.) *Text and technology. In honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, 233–250.

Berland, U. 1997. *Invariant tags: Pragmatic functions of innit, okay, right and yeah in London teenage conversations*. Unpublished MA thesis. Department of English, University of Bergen. Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., S. Conrad, & R. Reppen 1998. *Corpus linguistics. Investigating structure and use.* Cambridge: Cambridge University Press.

Bynes, A. 1998. *A corpus-based study of expletive use among London teenagers*. Unpublished MA thesis. Department of English, University of Bergen.

Chesterman, A. 1998. *Contrastive functional analysis*. Amsterdam/Philadelphia: John Benjamins.

Davidsen Bjerga, T. 1998. *Continuative and habitual aspect in English and Norwegian, with special reference to the English verb keep and the Norwegian verb pleie*. 'Hovedfag' thesis. Department of British and American Studies, University of Oslo.

Ebeling, J. Forthcoming. *Presentative constructions in English and Norwegian*. Dr.art. thesis. Oslo: Department of British and American Studies, University of Oslo.

Fabricius-Hansen, C. & K. Solfjeld. 1994. *Deutsche und norwegische Sachprosa im Vergleich. Ein Arbeitsbericht*. Arbeitsberichte des germanistischen Instituts der Universität Oslo Nr. 6. Oslo.

Filipovic, R. 1969. The choice of the corpus for a contrastive analysis of Serbo-Croatian and English. *The Yugoslav Serbo-Croatian – English Contrastive Project. B. Studies 1.* Zagreb: Institute of Linguistics, University of Zagreb.

Forskåhl, M. Engelska inslag i två finlandssvenska rollspel. Paper presented at the UNO workshop 4–6 September 1998 at Hanaholmen.

Fremer, M. *Va e du då* – generiskt *du* hos ungdomar och vuxna talare. Paper presented at the UNO workshop 4-6 September 1998 at Hanaholmen.

Gellerstam, M. 1996. Translations as a source for cross-linguistic studies. In K. Aijmer, B. Altenberg & M. Johansson (eds.) *Languages in contrast. Papers from a symposium on text based cross-linguistic studies, Lund 4–5 March 1994.* Lund: Lund University Press, 53–62.

Granger, S. 1998b. The computerized learner corpus: a versatile new source of data for SLA research. In S. Granger (ed.) *Learner English on computer*. London & New York: Addison Wesley Longman, 3–18.

Granger, S. & B. Altenberg. (Forthcoming.) The grammatical and lexical patterning of *make* in native and non-native student writing. Paper presented at the 3rd International Symposium on Phraseology, Stuttgart, 1–4 April 1998.

Granger, S. & S. Tyson 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English, *World Englishes* 15: 19–29.

Guldbæk-Ahvo, K. Brug af direkte anførsel i unge danskeres hverdagssamtaler. Paper presented at the UNO workshop 4–6 September 1998 at Hanaholmen.

Gunnarsdotter-Grönberg, A. Language, identity and lifestyle. Paper presented at the UNO workshop 4–6 September 1998 at Hanaholmen.

Hasselgren, A. 1999. *Smallwords and valid testing*. Unpublished PhD dissertation. Department of English, University of Bergen.

Hilmisdottir, H. & C. Wide. *sko* – en mångfunktionell diskurspartikel i isländskt ungdomsspråk. Paper presented at the UNO workshop 4–6 September 1998 at Hanaholmen.

Holmes, J. 1995. *Women, men and politeness*. London: Longman.

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

Johansson, S. 1998a. On the role of corpora in cross-linguistic research. In S. Johansson and S. Oksefjell (eds.) *Corpora and cross-linguistic research: theory, method, and case studies. Amsterdam: Rodopi, 1–24.*

Johansson, S. 1998b. Loving and hating in English and Norwegian. In D. Albrechtsen, B. Henriksen, I. M. Mees & E. Poulsen (eds.) *Perspectives on foreign and second language pedagogy*. Odense: Odense University Press, 93–103.

Johansson, S., J. Ebeling & S. Oksefjell. 1999*. The English-Norwegian parallel corpus: Manual*. Oslo: Department of British and American Studies, University of Oslo.

Kennedy, G. 1998. *An introduction to corpus linguistics.* London and New York: Longman.

Kotsinas, U-B. Forthcoming. Engelska i svensk slang. To appear in *Den unika Norden.* Published by Nordiska Ministerrådet.

Lauridsen, K. 1996. Text corpora and contrastive linguistics: Which type of corpus for which type of analysis? In K. Aijmer, B. Altenberg & M. Johansson (eds.) *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994. Lund: Lund University Press*, 63–71.

Lainio, J. 1997. Sverigefinska ungdomars slanguttryk. In U-B. Kotsinas, A-B. Stenström & A-M. Karlsson (eds.) *Ungdomsspråk i Norden. Stockholm: Meddelanden från Institutionen* för nordiska språk vid Stockholms universitet MINS 43, 188–203.

Løken, B. 1997. Expressing possibility in English and Norwegian, *ICAME Journal* 21, 43–57.

Monstad, K. 1998. *The grammaticalisation of sort of and kind of in young and old Londoners' speech*. Unpublished MA thesis. Department of English, University of Bergen.

Mosaker, H-M. 1998. *Qualifying utterances: A study of epistemic phrases in COLT and the BNC*. Unpublished MA thesis. Department of English, University of Bergen.

Møller, J. Identitetsaspekter ved sprogvalg hos tre tosprogede piger. Paper presented at the UNO workshop 4–6 September 1998 at Hanaholmen.

Odlin, T. 1989. *Language transfer: Cross-linguistic influence in language learning.*

Cambridge: Cambridge University Press.
Pedersen, A. 1995. *Vocabulary test development for Norwegian 8th graders*. Unpublished MA thesis. Department of English, University of Bergen.
Petch-Tyson, S. 1998. Reader/writer visibility in EFL persuasive writing. In S. Granger (ed.) *Learner English on computer*. London & New York: Addison Wesley Longman, 107–118.
Quist, P. Unge, identitet og sprog. Paper presented at the UNO workshop 4-6 September 1998 at Hanaholmen.
Ringbom, H. 1998. Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (ed.) *Learner English on computer*. London & New York: Addison Wesley Longman, 41–52.
Schiffrin, D. 1987. *Discourse markers*. Cambridge: Cambridge University Press.
Stenström, A-B. 1994. *An introduction to spoken interaction*. London: Longman.
Stenström, A-B. 1995. Taboos in teenage talk. In G. Melchers & B. Warren (eds.) *Studies in Anglistics*. Stockholm Almqvist & Wiksell International, 71–79.
Stenström, A-B. 1998. From sentence to discourse: *cos* (*because*) in teenage talk. In A. Jucker & Y. Ziv (eds.) *Discourse markers. Descriptions and theory*. Amsterdam:Benjamins, 127–46.
Stenström, A-B. Forthcoming. It's enough funny man; intentifiers in teenage talk. To appear in Proceedings from ICAME 19–98.
Stenström, A-B & G. Andersen. 1996. More trends in teenage talk: A corpus-based investigation of the discourse markers *cos* and *innit*. In C. Percy, Ch. Meyer & I. Lancashire (eds.) *Synchronic corpus linguistics.* Amsterdam: Rodopi, 189–201.
Strand, J. 1998. *Grammatical accuracy in English spoken by Norwegian 9th graders.* Unpublished MA thesis. Department of English, University of Bergen.
The Oxford English Dictionary. 1989. Oxford: Oxford University Press.
Tjerandsen, E. 1995. *Stories you may think is boring, can she make funny: Sentence complexity in the written work of Norwegian 8th graders*. Unpublished MA thesis. Department of English, University of Bergen.
Urdal, B. 1995. *Vocabulary in Norwegian 8th graders' written work*. Unpublished MA thesis. Department of English, University of Bergen.
Virtanen, T. 1998a. Direct questions in argumentative student writing. In S. Granger (ed.) *Learner English on computer*. London & New York: Addison Wesley Longman, 94–118.
Virtanen, T. 1998b. Argumentative uses of the progressive in NS and NNS student compositions: Notes on clause status and grounding. Paper presented at the International Symposium of Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, The Chinese University of Hong Kong, 14–16 December 1998. Extended abstract included in the symposium proceedings, 119–121.
Virtanen, T. Forthcoming. The Nordic subcorpora of the International Corpus of Learner English: A progress report. In S-K. Tanskanen & B. Wårvik (eds.) Proceedings from the 7th Nordic Conference for English Studies, Turku, 28–31 May 1998. University of Turku.
Virtanen, T. & S-A. Lindgrén 1998. British or American English? Investigating what EFL students say and what they do. In H. Lindquist, S. Klintborg, M. Levin & M. Estling (eds.) *The Major Varieties of English: Papers from MAVEN 97, Växjö 20–22 November 1997*. Växjö: Acta Wexionensia, Humaniora No.1, 273–281.
Wirdenäs, K. Diskurspartiklar i göteborgsgymnasisters samtal. Paper presented at the UNO workshop 4–6 September 1998 at Hanaholmen.