Harakka T. & M. Koskela (toim.) 1996. Kieli ja tietokone. AFinLAn vuosikirja 1996. Suomen soveltavan kielitieteen yhdistyksen (AFinLA) julkaisuja no. 54. Jyväskylä. s. 157 – 166.

EXPLOITING THE INTERNATIONAL CORPUS OF LEARNER ENGLISH (ICLE)

Tuija Virtanen

Åbo Akademi

ABSTRACT

The present paper deals with the *International Corpus of Learner English* (ICLE), currently under compilation and nearing completion. ICLE is a computerized corpus of argumentative EFL writing of advanced learners of various mother tongue backgrounds. Further, there is a comparable corpus produced by native speakers of English. In addition to research into non-nativeness, ICLE permits a large number of contrastive applications. Obvious fields to profit from this corpus include grammar and vocabulary but ICLE can also be used in discourse studies. The second part of the paper discusses some work already in progress where this corpus is being used.

Key words: *corpus, learner language, English*

0. INTRODUCTION

In this paper I shall first present the *International Corpus of Learner English (*ICLE) and then briefly discuss exploration or exploitation of this corpus for a number of different purposes.[1]

1. ICLE

ICLE is a computerized corpus of argumentative EFL essay writing by advanced learners of various mother tongue backgrounds. The ICLE project was launched in 1990 by Sylviane Granger at the University of Louvain-la-Neuve in Belgium. According to Granger (1993a:57), "the rationale behind creating a computerized corpus of learner English was to make use of advances in applied linguistics and computer technology to effect a thorough investigation of the interlanguage of the foreign language learner."

ICLE consists of a number of subcorpora, which represent learner English by university students of different mother tongue backgrounds: French, Dutch, German, Spanish, Czech, Polish, Japanese, Chinese, Russian, Swedish, and Finnish. The minimum size of each subcorpus is 200,000 running words, which adds up to over 2 million words for the entire corpus. Moreover, there is a comparable corpus produced by native speakers of English. In addition to research into non-nativeness, ICLE thus permits a large number of contrastive applications.

ICLE is currently under compilation and nearing completion. It is scheduled to be available in grammatically tagged form by 1997. The markup, tagging, and parsing of the data will be done in Louvain, with the Nijmegen tagger and parser, based on the Greenbaum and Quirk (1990) grammar. This is the TOSCA analysis system developed by Jan Aarts in Nijmegen and now applied to non-standard data in cooperation with the Louvain team (cf. Aarts 1995; Granger et al. 1994). There is a choice between an output consisting of non-disambiguated tags, i.e. all the possibilities listed one after the other, or disambiguated tags, which involves the first choice made by the program. Further disambiguating must be manual, which unfortunately makes it an improbable result in all the subcorpora. For information about the tagger and the parser in the analysis of interlanguage data, the reader is referred to the Louvain team: Sylviane Granger, and her students Fanny Meunier and Stephanie Tyson (cf. Granger et al. 1994; Meunier 1995).

Furthermore, there is an ongoing project in Louvain to develop an error tagger, which will enable all ICLE participants to code errors in the same way. This work is carried out by Sylviane Granger and Jonathon Guildford (1995; cf. also Granger et al. 1994), who are producing an error tagged Belgian learner corpus together with an error tagset manual (comparable to the ICE tagset manual, which it complements), and possibly a semi-automatic error tagging assistant.

ICLE will eventually be linked to the *International Corpus of English (ICE)*, led by Sidney Greenbaum and based at University College London, and it will thus be made commercially available. Software for data retrieval is currently being discussed. As *ICECUP (ICE Corpus Utility Program)*, developed in London, is not ready for use, other programs such as *TACT* and *WordCruncher* have so far been used to retrieve data from the corpus.

ICLE consists of argumentative essays, of 500 - 1000 words each, written by advanced students of English, mainly second-year or third-year students at departments of English at various universities. The topics are general and hence permit students to express themselves freely. The aim has been to collect the students' English as it is, rather than English influenced by their reading of reference books or articles in English. Also, students have been asked not to include quotations; the few that still appear in the essays will

be marked up as quotes and will thus not show up in frequency counts. Here are some examples of the kinds of topics students have been asked to write on; the idea has been to provide students with a statement to agree or disagree with, which allows them to produce an argumentative piece of writing.

TABLE1. Examples of topics figuring in ICLE.

Europe 1992: Loss of sovereignty or birth of a nation?
Money is the root of all evil
Crime does not pay
All armies should consist entirely of professional soldiers:

There is no value in a system of national service
Is violence a natural part of animal and human nature?
On what merits should the president of a country be elected?

Most of the essays have been untimed and students have then been able to use reference tools such as dictionaries, thesauruses, or grammar books. Further, some of the essays originate from timed exam situations, such as advanced proficiency exams. Finally, most subcorpora include a small portion of literature essays and exam papers, which may, however, turn out to be fairly different from the main body of argumentative data.

Each student is asked to fill in a special learner profile sheet, which gives background information about the essay writing situation, details about the student's and her/his parents' mother tongue, language(s) spoken at home, language(s) of instruction at school and university, years of English at school and university, stay in an English-speaking country, and other necessary particulars. Students sign the form to give permission for their essay to be used for research purposes. Learner profiles thus provide researchers with information which allows comparisons across subcorpora or sections of the corpus, selected for instance, on the basis of the mother tongue and/or second language, the language or languages spoken at home, or the language of instruction. All of this information will be coded in the final corpus.

At present the comparable native speaker corpus comprises some 95,000 words of student writing from Surrey, England, and some 68,000 words of native-speaker student writing from a number of American universities. This corpus will grow in size as it is of great importance for users of ICLE to be able to make comparisons between texts produced by non-professional NNS (non-native speaker) and NS (native speaker) writers, rather than compare NNS student writing with NS professional writing.

2. THE FINNISH SUBCORPUS

Before closing the presentation of the ICLE corpus, to go on to discuss its exploitation, let me say a few words about the Finnish subcorpus. This subcorpus is collected under the leadership of Håkan Ringbom, and it comprises essays written by Finnish-speaking and Swedish-speaking students. Some of the students are bilingual.

To start with, the Finland-Swedish materials primarily come from Åbo Akademi. As the Finland-Swedish students share the language with the Swedes and the educational background with the Finnish-speaking students, these data permit interesting comparisons between the Finnish subcorpus and the Swedish one, collected at the University of Lund. For the essays written by Finnish-speaking students we at Åbo Akademi are indebted to a number of English departments in this country. A large part of the existing materials come from the University of Jyväskylä, the rest mostly comprising argumentative essays written by students at the Universities of Joensuu and Helsinki, with a small literary sample from the University of Turku.[2] Our aim is to collect at least the minimum of 200,000 words from Finnish speakers. The Finland-Swedish corpus is smaller but we will obviously go on collecting material at Åbo Akademi, to compile a comparable Finland-Swedish corpus.

3. EXPLOITING ICLE

Once the entire ICLE corpus, with its eleven subcorpora, is commercially available on CD-ROM, it will thus be possible to make comparisons between the various subcorpora, to get at general aspects of non-nativeness. Moreover, it will be possible to use the American and British native-speaker corpora as a point of comparison. Some of the subcorpora come from bilingual countries and offer insights into research on bilingualism, while the various subcorpora will of course permit a large number of contrastive applications, to study cross-language and/or cross-cultural differences, with relevance to various kinds of professional writing that affect students and which these students may later engage in. Let me now turn to some of the work in progress in which this corpus is already being used.

An obvious field to profit from the quantitative results which can be obtained from ICLE with the help of various programs is the study of vocabulary. But ICLE differs from many other corpora of written language in that it also permits discourse applications: The texts that it consists of are entire discourses, and their discourse type and the situation in which they have been produced are, for most purposes, similar enough throughout the subcorpora.

At the University of Louvain, the exploration of the subcorpus representing the English of the French-speaking Belgians has mostly been focused on vocabulary and such discourse phenomena as are readily retrievable with the help of existing software. Hence, pedagogically interesting applications include the study of prepositions, idioms and collocations, word formation, and other lexical phenomena (e.g. lexical density, lexical sophistication, and 'foreign' words) - areas where we can really profit from these data. Furthermore, Sylviane Granger (1995) has worked on so-called 'false friends' in native-speaker (NS) and non-native-speaker (NNS) English, using the Belgian subcorpus and the British NS corpus of student writing. This is also an interest of Håkan Ringbom, who has investigated vocabulary in the essays written by Swedish-speaking and Finnish-speaking students in this country, comparing them with the Swedish and Belgian subcorpora, and the British NS corpus of comparable student writing. Ringbom shows for instance that many vague and stereotyped expressions are much more common in NNS writing as compared to the NS corpus. Another interesting difference is the underuse of low-frequency words and overuse of high-frequency words in NNS essays, as compared to NS student writing. In Ringbom's study, some of the core adjectives in NNS writing, such as *important* and *different*, are much less common in the British NS data. Further, high-frequency verbs in NNS writing include *think, get, make, become,* and *take*, which together with many other lexical items would seem to indicate a more informal style in the Finnish and Finland-Swedish data - a phenomenon also pointed out by Bengt Altenberg (1995) concerning the Swedish subcorpus. Finally, using an expression coined by Angela Hasselgren (1994), Ringbom concludes by singling out the 'lexical teddy bears' in the subcorpora under attention. Hence, the lexical teddy bear of the Finland-Swedish student seems to be the expression *I think*. This may be partly due to our emphasis of expressing a personal opinion in the extensive writing programme at our department, where students are asked to produce term papers from the very beginning of their studies. Secondly, a lexical teddy bear of the Finnish-speaking students, again, turns out to be *it is,* or *it's*, while an important expression in the Swedish data is *we must*.

To proceed now to the discourse studies where ICLE has been used, it is clear that the use of connectives, or connectors, is again very much in focus, due to easy retrievability. These studies concentrate on the general frequency of explicit linking signals in texts in various subcorpora, as compared to the native-speaker corpus, and further, the choice of connectors as shown by frequencies of individual linking signals. Hence, Sylviane Granger (1993b; 1994) and Stephanie Tyson (1995) have shown that individual conjuncts are overused or underused by NNS students, as compared to NS student writing. French-speaking students seem to overuse items such as *for instance, e.g. moreover, on the contrary, nevertheless,* and *namely*, and underuse some of the conjuncts favoured by British NS students, for example, *however, yet, instead, hence, therefore,* and *thus* (Granger 1993b; 1994; Tyson 1995). Similarly, Bengt

Altenberg (1995) points out that though Swedish students use conjuncts to the same extent as British NS students, there are differences in the choice of individual items. Thus, Swedish NNS writers overuse individual conjuncts such as *furthermore, for example, that is,* and *of course*, while an underuse is found of the following items: *therefore, thus, so, however*. A study of the Finnish use of connectors is yet to be done but Mauranen's (1993) analysis of academic papers raises expectations of a lower total of explicit conjuncts in this subcorpus, as compared to several other subcorpora or NS writing.

A quantitative study of connectors appearing in a body of data can obviously only be a start. As in discourse studies in general, a qualitative analysis of the overall structure of the individual texts is a necessity for reliable results. Explicit signals of a text strategy are not usually confined to a small, closed-class category of items, but involve a whole range of markers of various kinds. Moreover, to build up a plausible text world around the text (see e.g. Enkvist 1989), readers also make use of implicit information inferrable from the text and relevant encyclopedic knowledge. Further, it is clear that the topic of the essay should be taken into account in a study of connectors (as in many other fields, such as vocabulary; cf. e.g. the list of topical words separated in Ringbom's (1995) study). The discourse topic may affect the student's choice of text strategy, and this can be partly signalled by connectors. Finally, conjuncts are text-type sensitive, and despite argumentative topics, students may have chosen to produce an expository, descriptive, or narrative text (for a discussion of the advantage of introducing a two-level typology into the analysis of authentic texts, see Virtanen 1992). It will of course be possible to investigate all of this with the help of the complete ICLE.

Next, the analysis of connectors suggests other major issues concerning textual differences across ICLE subcorpora. But many textual and discoursal phenomena of interest are harder to get at with the help of existing software, and a manual analysis of the texts then seems the only possibility. Students, however, like to make comparisons across subcorpora in their term papers, and some of the discourse phenomena my students have looked at in a small scale include information structure, topic shifts, anaphora, hedging, attribution of knowledge to a source, and the expression of personal opinion. Bengt Altenberg recently (1995) reported on the results of his seminar in Lund, where students had studied various lexical phenomena, idioms, conjuncts, noun phrase complexity, nonfinite clauses, and involvement vs. detachment in the Swedish subcorpus and the British NS corpus.

It is to be expected that ICLE should manifest cultural differences in argumentative strategies across subcorpora and in comparison to NS student writing. The ICLE material consists of student writing on topics which invite argumentation. In other words, students are encouraged to take a stance for or against a problem. Further, some of these topics generate essays based

on a comparison between two points of view - a topic traditionally represented in Finnish matriculation exams (cf. Isaksson-Wikberg 1992). These data could thus profitably be used for a detailed study of the kind conducted by Isaksson-Wikberg (1992) on argumentative composition.

In her study based on materials which she had collected at Åbo Akademi, Isaksson-Wikberg showed that argumentative strategies in Swedish-speaking Finland differ from those found in American writing in several important respects (Isaksson-Wikberg 1992; cf. also her 1987 publications under the name Ingberg). Firstly, the placement and indeed presence, of a thesis statement, asserting the point of view chosen in the argumentative piece of writing seems to be the opposite: Americans are instructed to express a thesis statement early in the text, and indeed seem to do this to some extent. Finland-Swedish essays, again, often lack a clear thesis statement, or if there is one, it is found at the end of the essay. Secondly, the same goes for paragraph structures, with initial topic sentences in 'Anglo-American' writing - a huge concept, which nevertheless writing people seem to use - and a lack of topic sentences or one later in the paragraph in Finland-Swedish writing. Thirdly, we Finns seem to have been taught to show a balanced view of whatever we write about, highlighting several points of view, and often leaving the options open for the reader to choose from. Taking a stance should not result in underestimating the reader, i.e. it need not be rhetorically underlined to the extent that Americans seem to be instructed to do. In American writing, again, this may be considered impolite; a high degree of writer-responsibility is desirable in text production. Also, the reader seems to be looking for a stance and hence interpret argumentative patterns such as those found in Finland-Swedish writing as reflecting a lack of clear thought.

What we thus seem to have here are opposite argumentative strategies and views of what politeness towards one's readers should imply in practice. Anna Mauranen's results (1993), based on a comparison of rhetorical strategies in scientific articles written by Finnish and Anglo-American scholars, point to the same direction.

In an ongoing study of a number of ICLE subcorpora, I have noticed important differences in the frequency of direct questions. It seems that French-speaking Belgians and Finns, in particular Finland-Swedes, make more use of questions than American or British students (Virtanen 1995). Looking closer at Finland-Swedish student writing, which at the time comprised some 16,000 running words, I could distinguish two main discourse functions served by direct questions in these essays, i.e. topical questions, which introduce or shift topics and which are followed by an answer, and rhetorical questions. The results of my analysis expectedly confirm Isaksson-Wikberg's findings concerning topic sentences. In other words, questions functioning as topic sentences in these essays tend to appear late in the paragraph, followed by a discussion of the issue in subsequent paragraphs. Topical ques-

tions also appear later in the essay than similar instances in NS data. There seems to be a need to start both the entire essay and many of its paragraphs with some background leading to the topic at hand. Paragraph divisions are obviously based on a number of different criteria; yet they are often suggested by the content of the text (see Stark 1988), which makes paragraphing in a text worth paying attention to.

Comparing this sample of Finland-Swedish essays with some 19,000 running words of British student writing, I noticed that in addition to placing topical questions earlier in the text, native speakers also place rhetorical questions somewhat differently. In these NS essays rhetorical questions seem to be more evenly distributed across the text than in the Finland-Swedish data, where they tend not to appear in the first and last paragraphs of the essay. Moreover, while NS essays normally end with a conclusion, which can thus also include rhetorical questions, the last paragraph is usually very short in the Finland-Swedish and Finnish data and a conclusion is commonly missing. Finally, clusters of questions appear somewhat later in the essay than single questions, in both materials. A detailed analysis of rhetorical questions is currently in progress.

These are a few of the fairly conspicuous differences which appear at first sight. It is still too early to make educated guesses about the motivations behind the use of direct questions in argumentative NNS writing. It is, however, possible that the use of direct questions can be related to a more informal and/or involved style in NNS writing as compared to NS data. Another guess is that they indicate cultural differences in rhetorical patterns in use in the different subcorpora.

4. CONCLUSION

To conclude, I have touched on some work in progress where ICLE material is being used. The next step will be a symposium on learner corpora, organized by Sylviane Granger and Håkan Ringbom at the 1996 AILA congress in Jyväskylä. The main problem so far is the software, which limits the automatic or semi-automatic use of ICLE and other corpora. The situation is particularly acute in the study of discourse. But it is already clear that though small in size, the specialized corpus will be interesting for a number of purposes: It will provide data for contrastive studies on vocabulary, grammar, and discourse, and it can easily be used by undergraduate students in a seminar, or as data for more extensive theses. Moreover, ICLE permits research into non-nativeness in general. Finally, some of its uses will involve pedagogical applications for teachers and students, and it will be used to produce teaching materials based on comparisons of real data on learner language.

[1] I am grateful to Håkan Ringbom for comments on this paper.

[2] Due thanks go to Riikka Alanen, Anne Pitkänen-Huhta, and Kari Sajavaara at the University of Jyväskylä; Gregory Watson and Roy Goldblatt at the University of Joensuu, who have collected essays on several occasions; Mary Hatakka at Helsinki University for organising an essay-writing competition for this particular purpose, as did Kari Sajavaara in Jyväskylä; Outi Pickering, Lindsey Hair, and Bo Pettersson at Turku University, who have provided us with literature essays and exam papers. There are probably many others at these departments who have given us their time and energy to collect essays and learner profiles. We are grateful indeed to all these people for cooperation and we have high hopes of receiving yet another sample from some of them and from other Finnish universities.

REFERENCES

Aarts, J. 1995. "Generalized and customized annotation of corpora." Paper presented at an international symposium entitled "Exploiting Computer Learner Corpora", Louvain-la-Neuve, Belgium, 28 January, 1995.

Altenberg, B. 1995. "Exploring the Swedish subcorpus of ICLE." Paper presented at the Åbo symposium on the *International Corpus of Learner English*, 13 October, 1995.

Enkvist, N.E. 1989. "Connexity, interpretability, universes of discourse, and text worlds." In S. Allén (ed.) *Possible worlds in humanities, arts and sciences: proceedings of Nobel symposium 65*. Research in Text Theory 14. Berlin: de Gruyter, 162-186.

Granger, S. 1993a. "International Corpus of Learner English." In J. Aarts, P. de Haan & N. Oostdijk (eds.) *English language corpora: design, analysis and exploitation*. Amster dam: Rodopi, 57-69.

Granger, S. 1993b. "The role of L1 transfer in L2 production: evidence from a computer ized learner corpus." Paper presented at the 10th AILA Conference in Amsterdam, August 1993.

Granger, S. 1994. "The learner corpus: a revolution in applied linguistics." *English Today* 39 (10/3), 25-29.

Granger, S. 1995. "Romance words in English: from history to pedagogy." Paper presented at an international "Words" Symposium in Lund, 25-27 August, 1995.

Granger, S. & J. Guilford 1995. "Error categorization and error tagging." Working docu ment, Centre for English Corpus Linguistics, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Granger, S., Meunier, F. & S. Tyson 1994. "New insights into the learner lexicon: a pre liminary report from the International Corpus of Learner English." In L. Flowerdew & A.K.K. Tong (eds.) *Entering text*. The Language Centre of the The Hong Kong Uni versity of Science and Technology & the Department of English of the Guangzhou Institute of Foreign Languages, 102-113.

Greenbaum, S. & R. Quirk 1990. *A student's grammar of the English language*. London: Longman.

Hasselgren, A. 1994. "Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary." *International Journal of Applied Linguistics* 4 (2), 237-260.

Ingberg, M. 1987a. "Finland-Swedish paragraph patterns in EFL student compositions." In I. Lindblad & M. Ljung (eds.) *Proceedings from the third Nordic conference for Eng lish studies*. Stockholm: Almqvist & Wiksell International, 417-426.

Ingberg, M. 1987b. "The use of topic sentences Finland-Swedish student compositions." *Trondheim papers in applied linguistics (TRANS)* 4. University of Trondheim, Depart

ment of Applied Linguistics, 118-136.
Isaksson-Wikberg, M. 1992. *A cross-cultural study of American and Finland-Swedish rhetoric and argumentative composition, with special reference to EFL composition teaching*. Åbo Akademi. Unpublished Licentiate thesis.
Mauranen, A. 1993. *Cultural differences in academic rhetoric: a textlinguistic study*. Frankfurt: Peter Lang.
Meunier, F. 1995. "Tagging and parsing interlanguage." Paper presented at an interna tional symposium entitled "Exploiting Computer Learner Corpora", Louvain-la-Neuve, Belgium, 28 January, 1995.
Ringbom, H. 1995. "High-frequency words in the ICLE corpus." Paper presented at the EUROSLA workshop on Vocabulary Acquisition, Dublin, 11 September, 1995.
Stark, H.A. 1988. "What do paragraph markings do?" *Discourse Processes* 11 (3), 275-303.
Tyson, S. 1995. "Manual and automatic analysis of connectors: chalk or cheese?" Paper presented at an international symposium entitled "Exploiting Computer Learner Corpora", Louvain-la-Neuve, Belgium, 28 January, 1995.
Virtanen, T. 1992. "Issues of text typology: narrative - a 'basic' type of text?" *Text* 12 (2), 293-310.
Virtanen, T. 1995. "Discourse functions of questions in advanced EFL writing: exploiting the International Corpus of Learner English." Paper presented at the annual AAAL conference, Long Beach, California, 25-28 March, 1995.