Sylviane Granger, Estelle Dagneaux, and **Fanny Meunier** (eds.). *International Corpus of Learner English*. Version 1.1. Université catholique de Louvain: Centre for English Corpus Linguistics, 2002. Reviewed by **Erik Smitterberg**, Stockholm University.

The corpus-based study of learner English, from scientific and pedagogical perspectives, is an area of research that is attracting more and more scholarly interest, as evidenced by publications such as Granger, Hung, and Petch-Tyson (2002). By combining insights from Second Language Acquisition theory and English Language Teaching practice with a corpus linguistic methodology, researchers are able to describe interlanguage features and suggest implications for language teaching with greater confidence than has hitherto been possible.

Any area of corpus linguistics is necessarily dependent on available, reliable, and – preferably – comparable corpora that can serve as sources of data. Although a look at the corpora used by the scholars who contributed to Granger, Hung, and Petch-Tyson (2002) reveals that several learner corpora are currently being compiled in different parts of the world, few of these corpora appear to be publicly available as yet. In addition, some of the corpora chiefly contain specific types of learner English, such as ESP English, or English produced in an examination situation that may be more or less specific to the nation where the examination takes place. While all of these corpora appear to be reliable and valuable sources of data, there is still a need for learner corpora that are publicly available and comparable across several native languages. The publication of the International Corpus of Learner English (ICLE) is an important step forward in this regard.

The ICLE is stored on a CD-ROM, which contains a database of the corpus texts and the learner profiles. A license agreement and a handbook are also included. All page references in the present review are to the handbook, which has four sections: a description of the corpus, a user manual, a survey of the status of English in the countries of origin of the learners whose essays were sampled,[1] and a list of ICLE-based publications.

The ICLE contains about 2.5 million words of learner English; it consists of academic writing – mainly argumentative – produced by "university undergraduates in English (usually in their third or fourth year)" (p. 14). The corpus is divided into "eleven national subcorpora" (p. 27) of between 200,000 and 278,000 words each. Eleven native language backgrounds are represented, but there is no exact match between the backgrounds and the subcorpora: learners with a Swedish language background, for instance, are represented in both the

Finnish and Swedish subcorpora.[2] The term "national" is somewhat misleading regarding some subcorpora: for instance, the French subcorpus consists of essays written in Belgium (by native speakers of French), and the German subcorpus of essays written in Austria, Germany and Switzerland. This potential source of confusion is not serious, given the powerful selection tool that comes with the corpus texts (see below), but may still puzzle users, who will be faced with a list of countries to choose from that does not match the list of national subcorpora.

The learner profiles are stored in a database, and contain a great deal of information on each essay and essay-writer (see below). The profiles are linked to the texts by essay codes, which contain, among other things, a national code and an institution code (e.g. FIHE for Finnish, Helsinki University). The texts are in ASCII format, untagged, and contain no markup except for essay codes linking each text to its profile, and codes for deleted quotes, deleted bibliographical references, and illegible words. The text format is designed to work well with software tools for linguistic analysis such as WordSmith Tools.

After the corpus has been installed and the program started, the Query window, which "consists of two superimposed pages which can be moved to the foreground by clicking on the relevant header tab" (p. 54), appears on the screen. These two pages represent a major strength of the ICLE package: they contain about 20 variables (alphanumerical, numerical, alphabetical, or selection lists) according to which corpus users can select texts. The coverage is impressive: it is possible to select essays according to features of the essay (e.g. type, length, and production circumstances) as well as features of the learner (e.g. sex, country, native language, language at home, age, and years of English at school). The advantage of this coding scheme is that corpus users can design their own tailor-made subcorpora, which clearly helps to increase the validity and reliability of, for instance, comparisons across native languages. For example, Aijmer (2002: 73f.) emphasizes the importance of controlling for topic in research on modality in learner writing; the ICLE package enables users to select essays according to both type ("argumentative", "literary", or "other") and (words in) title. The only drawback in this respect is that some of the subcorpora that are selected by combining several variables will be quite small.[3] The handbook describes the selection process well, and help files are also available via the menu system of the program itself. However, some further information on how, exactly, each variable has been classified might be a useful addition to the handbook. Moreover, one variable I missed was whether each text had originally been submitted electronically or on paper (both methods were used), as this may affect the number and type of spelling errors. On the one hand, the use of spell

checkers may reduce the number of erroneous spellings; on the other hand, if spell checkers are not used, the keyboard also makes misspellings possible that would not be likely to occur in a handwritten essay (e.g. *bsd* for *bad* owing to the adjacency of the *s* and *a* keys, or *langauge* for *language* owing to fingers hitting keys in the wrong order).[4] However, the overall impression of the Query window is that of a very powerful tool indeed.

After carrying out the selection process in the Query window, the user is ready to click the "Search" button. This takes him/her to the Response window, where the search results are displayed in a grid, with the texts selected as rows and the variables as columns; there is also information on how many texts were selected. Among other things, the user can sort the essays according to their values on the variables (though only for one variable at a time), view, save, and print each text selected, and generate search reports that list the variables used and provide detailed profiles on each essay and essay-writer. However, the most important function may be the "Merge texts" option. This makes it possible for the user to conflate all texts selected into one single subcorpus, which "can then be printed or saved in an ASCII file for further processing or analysis" (p. 67). There are several reasons why this is a very useful feature. First, researchers can devote time to creating subcorpora that are comparable across several variables; they can then save these subcorpora as separate files and carry out several linguistic analyses on them without having to go through the selection process again. Secondly, the subcorpora selected can be processed further: for instance, part-of-speech tagging or error tagging could be supplied.[5] This feature is another major strength of the ICLE package.

The merged file can be saved on the researcher's hard disk and then analysed using text retrieval software tools. This process is very simple and straightforward. However, when I used WordSmith Tools to run a search for expressions of the future in Swedish essays, a potential problem appeared: a few expressions seemed to have the same context in the concordance.[6] Further investigation revealed that the two essays with the codes SWUG2028 and SWUG2040 were virtually identical; there were only a few differences regarding, for instance, word order (e.g. *rich, well-off people* vs. *well-off, rich people*), spelling (e.g. *mobil phones* vs. *mobilphones*), and punctuation. Several mistakes were also the same in the two essays (e.g. *looses out* for *loses out*), which suggests a common origin. This inclusion of virtually the same essay twice in the corpus need not be due to a mistake on the part of the compilers; instead, it may be the result of plagiarism, which is becoming a widespread problem in EFL composition courses.[7] Nonetheless, the discovery of two texts that are virtually identical in the corpus prompted me to look for further examples. Owing to time

limitations, I only carried out a few investigations in this regard, creating subcorpora and running searches to see whether the resulting concordance would reveal identical passages. The problem does not appear to be widespread, but an analysis of the occurrence of *might* in texts by German students revealed at least one other case of two essays that appeared identical.[8] On the one hand, these problems affect less than one per cent of all texts selected in the respective searches, and are thus unlikely to have any significant impact on quantitative results. On the other hand, there may be further identical texts that I have not discovered, as the search word(s) had to appear in the relevant essays for the inclusion of identical texts to be detected.

In sum, the publication of the ICLE is a milestone in the corpus-based study of learner English. The fact that researchers can easily create subcorpora of their own and the power of the software tool that allows them to do so are significant advantages. The long list of international collaborators makes it clear that a truly impressive coordinating effort must have been required to make all subcorpora comparable. It is to the editors' credit that they point out some limitations as regards the current version of the product, such as the lack of linguistic annotation and the fact that about 200,000 words per national subcorpus "precludes any investigation other than that of high frequency linguistic phenomena" (p. 38). The inclusion of a bibliography of ICLE-related publications, brief descriptions of learner corpus research methodology, and brief articles on the status of English in the countries of origin of the learners further adds to the usefulness of the publication. It is hoped that future versions of the ICLE will include tagged texts and further subcorpora (both of which the editors aim to do), as well as more details concerning the coding scheme for the ICLE database. Revisions of the database to ensure that identical essays do not occur in the material would also be welcome. Subsequent versions of the ICLE could thereby improve on the highly promising impression of version 1.1.

Notes

1. Austria and Switzerland, which account for a mere 70 and 60 essays respectively, are not included in the survey.
2. The national subcorpora – and native language backgrounds – present in version 1.1 of the ICLE are Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. Subsequent versions aim to include texts by Brazilian, Chinese, Japanese, Norwegian, Portuguese and South African learners also.

3. For instance, a search for argumentative essays written by male Spanish Spanish-speaking students who did not produce their essays in an examination situation yielded 15 essays of between 306 and 1,101 words in length.
4. One of the variables makes it possible for researchers to select only essays that were (or were not) produced with the use of reference tools, but as a reference tool may be both a dictionary for a handwritten essay and the spell checker of a word processor, this variable probably cannot be equated with that of whether the essays were submitted electronically or as handwritten documents.
5. However, the license agreement supplied in the handbook states that "[l]icensee shall not modify, decompile, disassemble, decrypt, extract or otherwise reverse the Product" (p. 49), and that users who wish to make other use of it are requested to contact the Licensor. In this respect, it is unclear to me what the legal status is as regards ASCII files that have been merged and saved separately: for instance, are licensees allowed to tag these merged files?
6. I am grateful to Petra Balog for originally drawing my attention to this issue.
7. However, most of the variables have the same values for both essays: for instance, they were written under exam conditions on the same day. This may suggest that the same essay was included twice, with different codes.
8. The two essays have the filenames GEAU3002 and GEAU3024 in the Response window. However, the essays linked to these codes appear to be identical, and both essays have the code <ICLE-GE-AUG-00024.3> in the Text window where the actual text file is presented, which may suggest an error in the coding scheme.

References

Aijmer, Karin. 2002. Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, and S. Petch-Tyson (eds.), 55–76.

Granger, Sylviane, Joseph Hung, and Stephanie Petch-Tyson (eds.). 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam and Philadelphia: John Benjamins.