Developing the Personal Dictionary

Ian Kemble

Ian Kemble is Deputy Head of the School of Languages and Area Studies at the University of Portsmouth. His main teaching subject is German, but increasingly his teaching reflects his scholarly interests in computer applications to language teaching and translation, on which he has published.
ian.kemble@port.ac.uk

Abstract
Apart from the dictionary, the computer has become the other essential tool of the translator. The article below is an account of a one-semester study unit (Computer Assisted Translation), available to undergraduate students in their final year, in which the computer is used to generate a glossary of technical terms, thereby replicating to some extent the experience of the modern lexicographer.

Background

When a group of foreign language teachers of the School of Languages and Area Studies first looked at the possibility of using computers in teaching in the mid-Eighties, two approaches presented themselves: either one went 'behind the screen' and learned a programming language (in those days it was a form of BASIC) in order to develop one's own teaching materials, or one stayed 'in front of the screen' and used commercially available, dedicated or authoring language learning software packages. I joined the group which opted for the latter, the end-user approach.

In many respects this perspective paved the way for the next important phase of our computer literacy development, the applications phase. This took the form of employing computer applications packages such as word processing in our teaching. For me the obvious application was to translation[1]. A resources book was developed for Year 2 students which took the form of 30 texts of progressive difficulty for translation from English into German[2]. Students would prepare for the discussion in class of the translated text by producing a draft version with the word processor in a small open access computer laboratory. The discussion of the text would then take place. The follow-on took the form of producing an improved version of the text, again with the word processor and, at the same time, a draft of the next text was prepared. The improved version was checked by the teacher. It was a simple idea, but it proved effective.

At the same time as this innovative class was launched I became increasingly aware of developments in lexicography involving corpora[3], and the idea of a course unit which combined translation with lexicography was born: the Computer Assisted Translation Unit, known as the CAT unit.

The CAT unit involves the students in two projects: a glossary and a machine translation project. In the following the focus will be on the former.

In outline, the first project involves students in developing a short **glossary** of 20 entries. Students identify a printed technical text of 5,000 words which becomes their electronically-stored **corpus**. The corpus is first submitted to a **concordancer** and then to a **dictionary generator**. The process of moving from a raw text or corpus to a glossary is analyzed in an **evaluative report** of approximately 2,500 words. The four main components of the project are examined in detail below.

The corpus

A corpus of 5,000 words is, of course, tiny. I can still remember digesting with some degree of incredulity the announcement that the Birmingham COBUILD project employed a corpus of some 20 million words. Nowadays, of course, large corpora consist of several hundred million words and there is talk of the first one billion word corpus (Landau 2001). However, our experience has shown that 5,000 words constitute an adequate corpus. Of course, small corpora present problems, just as large ones do. Words with a frequency of just one are quickly reached and, even with the limited objective of identifying just 20 glossary entries, some terms have to be selected from those with the lowest frequency. Students have to justify the selection, and inclusion in the glossary, of such low frequency terms.

As to the corpus itself, since these are foreign language students the corpus is in the foreign language, the glossary direction being foreign language > native language (L2-L1). In the first few years (the unit has been offered since 1990), there was an insistence that the corpus should be technical in the pure or applied scientific sense of the word. With time, a different perception has been reached, in which socio-scientific corpora of a more general nature are also accepted. Most texts are hybrids of common and technical language. It is pointed out to students that different text types generate different problems from the perspective of dictionary compilation. Providing students are aware of the issues and have some idea of how the problems can be tackled, they are usually in a position to resolve the remaining issues which may occur, and to justify their decisions.

The School of Languages and Area Studies at the University of Portsmouth is one of the largest departments in the UK teaching five languages (English, French, German, Italian and Spanish) and associated studies to more than 900 students on over 20 courses.
www.hum.port.ac.uk/slas/main.htm

Concordancer

Initially, Micro-OCP (Oxford Concordance Package) was used. With the advent of Windows a package, compatible with the new environment, has been identified, namely MonoConc[4]. The basic features of the concordancer programme are utilized: frequency lists, the standard KWIC (keyword in context) concordance, more sophisticated concordances, such as left- and right-sorted concordances. The majority of students approach the task in a reductionist way, and are, perhaps, not always willing to experiment with the facilities of the package in a dynamic way, but there are always the minority who are more adventurous and realize that concordances can be quickly produced and analyzed.

Hitherto, no statistical measures have been included in the project, such as the lexical density measure or the type:token measure, but that is set to change (see section on The Future below).

The procedure is the normal one, initially to produce two frequency lists: a first based on frequency of occurrence and a second based on alphabetical listing. With such small corpora the alphabetical list comes into its own, grouping together words with the same stem and thereby creating a potential for glossary entries to be identified. The project guidelines require students to comment in the evaluative report on the way the two lists complement each other in terms of the information they provide for the lexicographer.

Secondly, a variety of concordances is produced which allows the identification of suitable multi-word terms for an entry into the glossary. The patterning in language we know as collocation has been described in terms of a cline (Carter 1987, Chapter 3, p. 63 in particular) with relatively loose patterning at one end of the spectrum (unrestricted collocation) and much more fixed patterning at the other (restricted collocations). Small corpora of 5,000 words are likely to produce much looser patterns, possibly ones which are unique to the text in question. However, their inclusion in the glossary can be justified in terms of the frequency of occurrence in the corpus in question. Students, on the other hand, veer away from relatively unresrticted collocations, despite their legitimate inclusion in the glossary, and prefer to be influenced – understandably, though unjustifiably – by the orthodoxy of the general dictionary. The concordancer produces not only single word terms and multi-word terms, but sample contexts in which the terms in question are used. This is particularly useful for the next phase, namely the compilation of the glossary. All information (e.g. frequency lists and concordances) is stored as separate files on a disc, supported by hard copy notes.

Dictionary generator

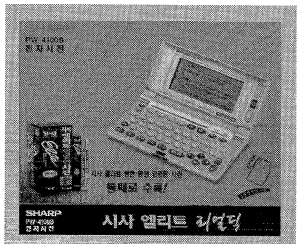
Again, the original software has been replaced by Windows software. Currently, TRADOS 95 Multiterm software is used which, among other things, has the advantage of being an industry standard.

TRADOS is a flexible package allowing the user to define the precise nature of the entry. In brief, there are three types of attribute, from which the user selects the features which are appropriate to the 'shape' of her or his glossary entry. They are index fields, which allow the languages to be used in the glossary to be specified; index fields which contain, for example, definitions, source information or notes; and, attribute fields, which contain information which can be classified, e.g. gender, subject field, etc.

The package is designed primarily for single workstation use and does not appear to operate completely free of technical intervention in a network environment. But the problems, small in number, are easily resolved by the technical support staff. Students have the task of defining the user of the glossary and designing their glossary accordingly. At the same time, they are required to include a number of definitions and uses of the term in context. Examples of the use in context are provided by the concordancer, the glossary definition has to be researched by the student, as does the equivalent of the term in the target language. In this way, a balance has to be achieved between flexibility (the student decides on the glossary entry) and rigour (the student justifies the glossary entry).

The glossary

The glossary originally contained 40 terms, but that has been progressively reduced to 20. It has to be borne in mind that the task is completed within six weeks, one half of the taught semester. In feedback sessions in the past, students have commented on the disproportionate amount of time devoted to the project, hence it has been made manageable and has, thereby, created the potential for projects of a higher quality. The glossary contains single terms, which are usually selected on the basis of a number of criteria which typically include: frequency of occurrence, level of technicality, keyword status, and multi-word terms which are provided by an analysis of the concordances which are generated.



SHARP PW-4100S
Sharp Korea
Seoul, 2001
Electronic Hand-Held Device containing the SI-SA ELITE English-English-Korean, English-Korean and Korean-English Dictionaries, with TOEIC material (by YBM Si-sa)
http://kdictionaries.com/products/medium-elect/med-el-korean.html

Evaluative report

The project evaluation provides the key to the project. It takes the form of a report in which the student reflects on the process of developing a glossary from a raw corpus. A justification for the decision taken at each stage is provided: selection of corpus, analysis of frequency lists, analysis of the concordances, determination of the shape of the glossary, construction of the glossary entries, etc. Its analytical content is a significant criterion in the marking of the project.

The unit has two other characteristics which are worthy of mention: **delivery** and **skills development**.

Delivery

The unit is delivered in a very different way to the standard practice of lecture and seminar, since technology is involved. All sessions take place in the computer laboratory. Information sessions precede workshop sessions. In the former, lexicography is introduced with particular reference to modern developments involving the computer, in the latter – which are more frequent – students work on their project with support from the teacher. Learning is experiential.

Skills development

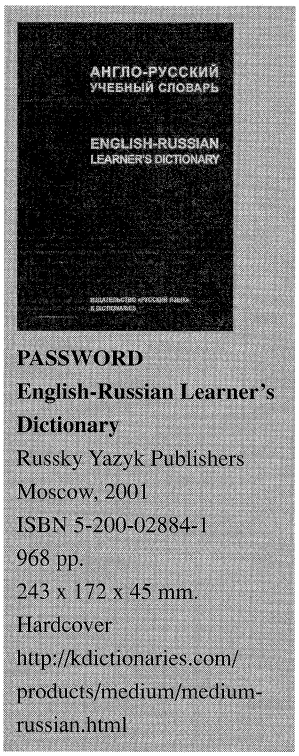
The unit has the advantage of assisting students with the development of skills, particularly those of analysis, problem-solving and IT. In the UK, evidence of key skills development is an important item on the agenda of the quality assurance agencies.

The Future

The unit has proved popular with students who are attracted by its vocational orientation and its potential to develop their IT skills. The School has recently embarked on a research project, the aim of which is to identify what students actually do in terms of vocabulary acquisition and development. The major finding is that only a third of the total student cohort sees itself as effective when it comes to vocabulary learning. As a consequence, a series of measures is to be taken to remedy the situation. These include the production by the School of a 'Guide to Vocabulary Learning' and the introduction into the Year 1 Study Skills programme of a session on 'Making Effective Use of the Dictionary'. The approach which is being adopted is characterized by a desire to achieve a balance between flexibility (allowing learners to develop their own vocabulary learning strategies) and rigour (insistence on the development of a personal dictionary). In the future, students will enter the final year of their programmes with an understanding of the principles of vocabulary organisation and have evidence to show for it in the form of a personal dictionary in the traditional printed format. The CAT option will then enable them to develop their own personal dictionary in electronic format.

Conclusion

We are not in the business of training lexicographers, but we are, and will be increasingly, able to provide students with both an understanding and experience of vocabulary across a range of contexts, including the organisation of vocabulary in a personal dictionary.



PASSWORD
English-Russian Learner's Dictionary
Russky Yazyk Publishers
Moscow, 2001
ISBN 5-200-02884-1
968 pp.
243 x 172 x 45 mm.
Hardcover
http://kdictionaries.com/products/medium/medium-russian.html

Notes

1. Interestingly, translation both as a vehicle for developing language skills and as an end in itself has survived the 'communicative revolution' which has characterized language teaching methodology in the UK for the past two decades. Indeed, translation has not only survived but has thrived.
2. *Translating for Pleasure* was first produced in 1987 and was followed by a second edition ten years later. Despite the aging of a number of the texts it continues to sell in the market place. This is evidence that there is a definite niche for such a resource.
3. The Collins COBUILD dictionary, first published in 1987 (now in its third edition, 2001, HarperCollins, Glasgow), was the pioneering dictionary of the new generation of dictionaries in the UK.
4. MonoConc for Windows, developed by Michael Barlow, is published by Athelstan (http://athel.com).

Short Indicative Bibliography

Brierley, W. and I. Kemble. 1991. *Computers as a Tool in Language Teaching.* Chichester: Ellis Horwood.
Carter, R. 1987. *Vocabulary.* London: Unwin Hyman.
Kemble, I. 1996. *Translating for Pleasure.* Portsmouth: Hampshire Open Learning Unit.
Landau, S. I. 2001. *Dictionaries, The Art and Craft of Lexicography.* (Second Edition.) Cambridge: Cambridge University Press.