11.3 "Compiling a Bilingual Dictionary in the Electronic Age."

Talk given by Ms. Sue Atkins of the Oxford University Press on 7 May 1992 at King's College, London.

Ms. Atkins began by outlining the work of a lexicographer in the compilation of dictionaries.

In general terms the analysis of words to be included is followed by the synthesis of entries bearing in mind the needs of the User of each version of the dictionary.

In a bilingual dictionary the complexity is increased because the entries in both sections of the dictionary have to be written with both source language and target language speakers in mind. It is interesting to note that truly bilingual speakers are not necessarily suited to this aspect of dictionary writing. It is generally necessary to employ both source language and target language editors who supply appropriate data as well as reviewing each others contributions.

Ms. Atkins then talked about the dictionary project to which she is a consultant, the new Oxford-Hachette English/French/English dictionary.

The processes for producing the dictionary have to be carefully controlled by the use of standard procedures. As an interesting sidelight Ms. Atkins said that it was found that access to other existing dictionaries had to be restricted to avoid delay and to ensure a new and independent view.

The following is a copy of the Project Description kindly provided by Ms. Atkins at the meeting.

Project Description

Overview

This project starts from the independent analyses of both source languages using standard documents ('frameworks'), in the form of lexical entries showing the use of each meaning in context and structured so as to be consistent and compatible; these are then translated, and from the translated lexical entries, a pair of lexicographers (one partner anglophone, one francophone, each highly proficient in the other language) tailor out the dictionary entries; the text is keyed, and a final editing read made.

Back-up expertise includes computational support, a team of terminologists for each language, a team of American English editors, and full project management, keyboarding, and secretarial staff.

The project requires approximately 170 person/years and represents an investment of 2.5 million pounds for the publishers involved. Since the price of the dictionaries is controlled by market forces, it is not expected to show a profit in the first ten years of the published dictionary's life.

Sources

These are entirely electronic and exist independently for the two languages: two 10-million word general language corpora and two smaller neologisms corpora, with interactive concordancing facilities offering both KWIC and sentence-length concordances; on-line dictionaries from both publishers; and an on-line terminology database. These resources are routinely used by all in-house editors.

Also selectively available are translation-equivalent sentence pairs extracted automatically from a bilingual corpus of parallel texts; these are mainly used to help with real equivalence problems.

Theoretical infrastructure

The lexicography is based on a linguistic analysis drawing on current theoretical work. Over one hundred 'template' entries in each language ensure consistency of approach to members of lexical sets (e.g. days of the week, colours, foods, etc); the work on English verbs is based on a theoretical analysis of the English verb system [Levin forthcoming]; the French verb entries benefit from the work of the LADL (Laboratoire d'Automatique Documentaire et Linguistique) team [Gross 1981 etc.]. The English and French style guides occupy 300 single-spaced pages and continue to grow.

Methodology

The 'framework' sheets for each word are prepared according to detailed instructions by native-speakers of the source language. These source-language frameworks constitute rich relational databases, quite independent one from the other. These are handwritten; the frameworkers are freelance compilers working out of house.

The framework entries are translated by native-speakers of the target language; handwritten, by freelance out-of-house translators, working according to rigorous guidelines. Frameworkers and translators mark some items for specialist treatment (from terminologists or American editors).

The translated frameworks, handwritten, go to the editor-pairs, one English and one French native-speaker, working in house. They compile the dictionary text, manually editing the translated framework, or rewriting it. It is a point of principle that English is written only by English native-speakers, and French by French native-speakers.

The text is then keyed into a database using the Standard Generalised Mark-up Language (SGML) with a tagset of over 50 tags. The keyed text goes to the American editors, and on its return is re-edited by in-house staff in order to incorporate amendments from these and other specialists. It is checked by the senior editors and the amendments are keyboarded.

The finished dictionary text is approximately one-third of the size of the compiled bilingual data (on frameworks). This complex editing process is not amenable to on-line compiling, given the computational resources of even the largest publishing house: many dictionary entries (far less translated framework entries) are too long to allow an overview on one screen.

Lexicographical talents are rare, keyboarding skills are easier to find. It is a waste of lexicographers' time to have them keyboard and tag text. It should be noted in passing that the publishers cannot afford to key, post hoc, the frameworks, translated or not, which are of no use to the current project and do not justify further expenditure.

Computational aspects

The computer has three roles in this project:

- first, it supplies the lexicographical evidence (from corpora etc) on which the dictionary text is based;
- second, it holds the final tagged text of the dictionary and controls the length of the compiled text according to pre-established targets;
- third, it allows this complex project to be planned and controlled efficiently, by means of a 'tracking system' built by OUP Reference Computing. The data in the tracking system consists of the approximately 60,000 headwords in each language, grouped by morphological sets, and graded according to lexicographic complexity. The system builds 'workpacks' for frameworkers, translators and editors according to specifications of size, quantity and complexity set by the project managers, and tracks each lexical entry throughout the life of the project. This allows the managers to control the project flexibly and efficiently.

Lexicographers

Prerequisite qualifications are the equivalent of a good honours degree in the other language, and two or more years recently spent in that linguistic community. The compiling team is highly qualified and the in-house bilingual lexicographers undergo full-time training for the initial period of their employment. It takes a year before an editor is fully trained.

References

Gross, Maurice (1981), "Les bases empiriques de la notion de prédicat sémantique; Formes syntaxiques et prédicats sémantiques", eds. A.Guillet & C.Leclère, "Langages", No. 63, Paris: Larousse, 7-52.

Levin,B. (forthcoming) "English Verb Classes and Alternations: A Preliminary Investigation", University of Chicago Press, Chicago, USA.