Bitext makes progress with and without the name

The term *bitext* was introduced in an article by *Brian Harris* of the University of Ottawa which appeared in *Language Monthly* in March, 1988. In its practical aspect it refers to 'aligning' (or 'keying') existing translations interlinearly or in parallel with the original texts in such a way that when any segment of a source text is retrieved the translation is retrieved with it. The texts are stored that way in a text bank in the computer. Then we can compare new source documents automatically with the texts in the text bank, and the previous translations will be found and displayed for any segments that are recurrences. The finds can then either be transferred just as they are to the new translations, or if they do not entirely fit the new contexts they may at least stimulate the translator's mind by analogies.

Everybody involved in translation knows that a great deal of time and effort is being wasted by re-translation of texts or pieces of text that have been translated before, but precisely how much? IBM European Language Services set out to answer this question for some of its own work. At its Bikerød (Denmark) facility, a student from Copenhagen University did research in the summer of 1988 on the amount of repetition in IBM manuals, both from one version of a product to the next, and between the manual for one product and the manual for another. The averages were found to be 15% repetition within a manual, 50% across manuals (*Language International* 1(6):6-7). Clearly the potential savings would justify the writing of efficient comparison-and-retrieval software.

A further implication was that besides the benefit from comparing new texts with old there is an additional (though smaller) saving to be made, as a translation progresses, by comparing each further passage of the document with the part already translated.

The saving might be even higher with some other types of document. For example, the translation service of Confederation Life Insurance in Montreal was already doing its best to avoid re-inventing what it calls 'standard translations', that is to say paragraphs that are repeated with little or no change from one insurance document to another. However, it was proving a tedious and imperfect process to look up these repetitions by hand.

In an update on bitext in December 1988, Harris pointed out that several of the components needed for bitext systems were already on the market. An interlinear word processor called IT had been available for some time from the Summer School of Linguistics in the United States. The FILIUS module of the Logos CAT system permitted simultaneous viewing of source and target texts even if stored separately:

> FILIUS automatically positions source and target files in two windows on the PC; when the translator proceeds to a new sentence in the target document, FILIUS automatically highlights **the corresponding sentence** in the source document. We call this 'synchronized scrolling'. (Jon Cave at the 1988 ATA Convention)

Likewise Xerox's VIEWPOINT displayed 'side-by-side' translations as feedback for improving the company's MCE (Multinational Customised English); sentences were paired and numbered. (*Maria Russo* at the same convention.)

It was therefore just a matter of putting the components together and compiling the text banks. Today complete custom-designed systems have made their appearance both in Europe and North America.

The first of these would seem to have been IBM European Language Service's Translation Support Facility (TSF), which, under the influence of the research mentioned above, incorporated a repeated sentence identification facility. Bitext had also been under development at Gigatext's ill-fated project in Saskatchewan, Canada, when the firm collapsed; so the first working

Brian Harris

system in North America is probably *John Chandioux's* GENERAL TAO for Confederation Life of Montreal (*Général Tao* is a pun on the name of a Szechuan dish served in Chinese restaurants and the acronym TAO for 'traduction assistée par ordinateur').

The developers of these systems are still feeling their way towards their full potential. A problem that soon became apparent with TSF was that writers of successive manuals or versions often used a slightly variant wording to convey the same message. For example one manual may contain an instruction, 'Type the text you want to center,' whereas another phrases it, 'Type the text you want centered'. Similar problems arise in the Confederation Life texts. The solution in both systems has been to introduce matching that is more flexible than completely verbatim replication; IBM calls it *fuzzymatch*. The question then is how fuzzy can the matches be without creating a lot of 'noise' (irrelevant retrievals)? Probably the future will produce 'conceptual matching' as well as verbal matching, and where several matches are found they will be ranked in order of relevance to the translation in hand.

TSF and GENERAL TAO are aids for human translators and not for machine translation. In the final analysis, therefore, it is the translator who decides whether a previous translation can be recycled in another context.

Bitext has been introduced into MT research too, but there it has taken unexpected and quite opposite directions. On the one hand it is at the heart of the IBM Thomas J. Watson Research Center attempt to *predict* translations by statistical methods. Such statistics require sampling large quantities of previous translations; hence a data base is used which is surely the biggest bitext anywhere — 100 million words of Canadian parliamentary proceedings, *Hansard*, in aligned English and French. However the alignment is only at the sentence level instead of at the more desirable clause and phrase level (*Computational Linguistics* 16(2):79-85.) In contrast bitext is also the basis of a small prototype Bilingual Knowledge Bank (BKB) of 2,500 sentences that is being experimented with at BSO Utrecht:

> A BKB is a structured parallel corpus of bilingual text. Its purpose is to serve as the primary source of linguistic and extra-linguistic knowledge for all of the modules involved in the machine translation process. (Victor Sadler, *Working With Analogical Semantics*, 1990.

Maybe we ought to call the BSO corpus 'tritext', since it is in English, French and Esperanto!)

Sadler points out that one pressing reason for drawing knowledge from a text corpus is the very high cost of compiling language processing system dictionaries by conventional means:

> Forty person-years' hard work has resulted in several thousand entries in each bilingual dictionary [of BSO's DLT prototype] ... A production system would need far larger sources, and these would have to be multiplied by the number of languages and the subject fields to be covered. How could this be achieved on any reasonable time-scale?

The compilation and upkeep of term banks for human translators is also very costly. Bitext is one way to attack the problem: once the texts are in the computer they can resource an instant bilingual contextual dictionary on demand for any word or phrase.

The concept of bitext has made more progress than the word itself. In fact of the developers mentioned above only Sadler uses the term. Nevertheless it has acquired two translations in French: 'bitexte' (*Christian Boitet*, France) and 'banque de données bitextuelles' (*Claude Bédard*, Canada). Whatever name this rose is eventually known by, it is already blooming.