[From: *Translating and the Computer*, B.M. Snell (ed.), North-Holland Publishing Company, 1979]

EURODICAUTOM

J. GOETSCHALCKX

Head of the Terminology Bureau
Medium- and Long-term Translation Service
Commission of the European Communities
Luxembourg

The original name for what we now so proudly call EURODICAUTOM was in fact DICAUTOM, standing simply for dictionnaire automatique.

This reveals that from the beginning, 15 years ago, when my predecessor Mr. J.A. BACHRACH started this project, we did not intend to set up a clever, sophisticated gadget. We wanted to offer the translator a new, more efficient tool, so that he could do his work as he did before, consulting a dictionary to resolve terminological problems if necessary.

EURODICAUTOM is not an automatic device, at least not a penny-in-the-slot-machine. It is just a computer-aided dictionary.

But it is of course a dictionary of a special type able to supply information more rapidly and more comprehensively than would a normal lexicographical work.

What kind of information can be offered by EURODICAUTOM?

First of all it can give corresponding phrases or sentences in several languages.

Secondly it can give terms or expressions called "vedettes", accompanied by illustrative but not corresponding contexts and/or definitions.

Finally it can give term-to-term equivalents. This unfortunately is the form in which most multilingual lexicographical work is produced.

The phraseological approach is very important in the scientific and technical field where we have to deal with so-called special languages. Specialisation by our translators is not always possible. For this reason we have to offer full information on the use of the terms in this particular field and very often also a lot of purely technical information. If these phrases or sentences are well chosen they can cater for both the linguistic and technical information needed.

For subtle distinctions between scientific or technical terms a definition is very often the best solution, but it is not always easy to find definitions for all terms appearing in a translation.

Although mere term-to term equivalents are often dangerous in the hands of a "multi-purpose" translator, we did not want to be cut off from this important source of scientific and technical terminology. We believe that being able to accept terminological information in any form of presentation is an ideal situation for exchanges with other centres.

What is the origin of the information we put into a terminology bank? Our own terminological information is the result of an analysis of original documents in each

language. The comparative study of these original documents gives real equivalents of professional language usage. This means that the equivalency is very often at the level of the phrase and not necessarily between corresponding words. The semantic content of the corresponding phrases is the same but it can be distributed in a different way from one language to another among the morphemes

But we also use terminology compiled by specialized institutions such as AFNOR, the French standards institute, the International Welding Institute, and the European Brewery Convention which is concerned not only about the quality of beer but also about its terminology.

In what order is this information presented? In our system each information unit is called an entity. I understand that in recent congresses on applied linguistics the French word "fiche" has been generally and internationally accepted for the same concept.

A "fiche" in automated terminology contains purely terminological information, documentation data, and information needed for the electronic data processing and general organisation of the data base and the whole system.

In the EURODICAUTOM system each "fiche" is determined by three elements: NI: identification number (as we started from French, the letters are the wrong way round, of course); BE: "bureau émetteur", which is the terminology office which created this terminological information or assumed the responsibility for its input; TY: "type", which is a three-letter code describing an homogeneous collection of terminology or a collection of individual cards designated by the code TFI (fiche individuelle, fiche isolée). This is what we call the "BETYNI" of a fiche.

These three elements determine exactly the terminology unit, the "fiche". This apparently unimportant detail has important consequences, It means that a network of terminology centres can feed EURODICAUTOM and, independently of one another, can use their own organisation, as far as choice of subjects for their terminology collections and their own numbering systems are concerned. There is no risk that it might interfere with the collections of other centres belonging to the network.

In practice this means decentralisation of information collection, centralised storage and distribution. This is important for, as you may know, the European Communities have several institutions each with their own translation department and usually a terminology office.

When we further examine this EURODICAUTOM fiche we see that it also mentions the author. We respect intellectual property. Nevertheless this possibility has not been used so far. Terminologists are modest people. Or could it be that they are not quite sure of the overall value of their inventions? In most cases we stick to professional usage. Our creative function in terminology is only called upon when there is no other way out.

Then we have a reliability code which in fact, has nothing to do with reliability because it is not a real measure of the reliability of the information. Terminologists can also have terminology problems of their own.

The so-called reliability code also indicates whether the information has the value of a standard. This is the case of international or national standardized terminology.

Extracts from the European Treaties or regulations are of the same type. In

quotations they have of course, to be reproduced literally as they stand in the official texts. This type of fiche bears the code 5.

If all the language versions are supported by solid sources, the fiche will be given the code 4. As soon as perfection is no longer guaranteed the figure is reduced.

This does not mean that the information is bad. We simply are not sure, for lack of bibliographical information. Consequently we know that we still have to work on this fiche to bring it up to the standards of good, reliable terminological information. So it still has something to do with reliability.

Scientific, technical, political and economic terminology, as well as a certain dose of EEC home-jargon represents an enormous mass of linguistic information. Some classification would therefore do no harm. This is why my colleague Dr. LENNOCH developed on the basis of the UDC a three-digit alphanumeric subject code system. It is now used not only for EURODICAUTOM but also for the Target-project of the Carnegie Mellon University in Pittsburgh, USA and by the terminology office of the World Bank in Washington. The multiple coding in this system results in a high degree of refinement.

Subject codes give a supplementary retrieval parameter for polysemic terms. They are especially useful for extracting miniglossaries on certain special fields from the corpus of the terminology bank. This material can serve as a valuable aide-me"moire for interpreters having to prepare for a meeting of experts on some exotic new technique.

All these documentary and data processing data are there to highlight and support the real heart of the matter, the purely terminological information in "vedette"or in context, with or without definition.

If there is something more to say about the "vedette" that cannot be part of a definition, such as nationally or geographically limited usage, peculiar plural form etc., this can be done in a scope note NT.

Until now I have only spoken about the organisation of the fiche, the presentation of the terminological unit with its explanatory documentation. Let us now come to the retrieval stage.

As I said in the beginning, EURODICAUTOM should offer the translator a working tool which allows him to achieve higher efficiency and quality without implying fundamental changes in his working methods .

After an extremely short and simple sign-on procedure, the terminal invites the user: "Type your question".

Although we have only 130,000 "fiches" or entities at the moment, we have taken into account from the beginning the necessity of having many more, 1 million perhaps, and the difficulties arising then because of polysemy etc.

That is why we have incorporated a weighting system. The idea was to give first of ail the best answer according to the principle of the longest match.

if a multiterm ABC is the subject of a question, the system first gives ABC if it is in the corpus, and then AB, BC or AC. This "partial" information can be useful. If not the translator just stops the interrogation. This reveals again our basic concept: a working tool for a specialist not a "penny-in-the-slot machine" for anybody.

Let me give an example: A translator is looking tor the translation of the technical expression "relative cinematic viscosity". The first answer gives the translation of this multiterm. But if the user continues with the interrogation he will obtain consecutively "relative viscosity" and "cinematic viscosity".

If the full expression had not been available, it would have been very easy to reconstruct it from these two "partial answers".

It is perfectly clear that the higher the number of terms in the expression looked up the more a partial answer is likely to give useful information. With two terms the risk of irrelevant information is much greater, but as as the system is made for translators they must be capable of judging immediately if the partial information for is useful or not.

To improve the system we shall reduce the partial answers to those containing not less than n-2 of the terms contained in the question.

Furthermore the partial answers to the question AB (two terms) will give alternatively answers with A only and answers with B only. Let us assume that you are looking for the translation of roll-on-roll-off-ship . If the system gives you a series of partial answers with the term "ship" it is highly unlikely that this will prove useful information. On the other hand any partial answer with the term "roll-on-roll-off" will give a useful hint for the right translation of the original expression. So to avoid a long series of poor information containing the term ship, the system will give alternatively both elements of the expression . "Roll-on-roll-off craft" e.g. would be helpful for translating the original "RO/RO ship".

Congress attenders of the clever type will have realised that there is a retrieval problem with the phraseological entries, because the words in the phrases are not always in the standard form. This is especially true of languages like German with its numerous inflections and Danish because of the suffixation of the article. To solve this problem we use the truncation device. If, for example, a phrase contains a plural form, truncation will still allow the information to be obtained. Even in Italian it provides the possibility of asking for a form ending in CA or CO and obtaining as an answer the plural form ending in CHE or CHI.

We can do even better: the expression "in and outgoing ships" is a form which is not very frequent in English but more so in German and in Dutch. A fiche containing the expression "Stuetz- und Bewegungsapparat" can be the answer to a question requesting the translation of Stuetzapparat.

For interrogation regarding a polysemic term or a document concerning a very specific subject field, the interrogation can be made after introducing one or more subject codes.

This should not eliminate other information corresponding to the question asked but without the subject code asked for.

Coding is often a very subjective matter but wouldn't it be a pity to lose information because of a mere coding error? On the other hand some terms can be common to different fields and have the same equivalent in other languages. Forming techniques in plastics are partly the same as in metals. A terminologist introducing this information on the basis of a document on metal forming could forget to also assign it the general code for mechanical treatment. A user asking about a document dealing with plastics forming might make the same mistake while composing his interrogation parameters.

As you can see, the system is almost foolproof.

Nevertheless there is still a lot to be done.

We still have to eliminate many redundancies and typing errors. Poor term-to-term information has to be enriched by contexts or definitions. Dutch, Italian and Danish are not as well represented as German, English and French. They have to be added.

Furthermore, 130,000 terms are not enough, 300 or 400,000 would be a better working basis. It is a matter of time and staff.

We hope that we shall obtain the necessary means and that exchanges with other terminology centres will be possible and helpful.

The creation of national terminology centers as is encouraged by UNESCO's Nairobi Recommendation on the setting up of terminology centers for special subjects or certain industrial industrial sectors would help the quicker collection of terms and better terminology work.

But, as I learned in my first Italian lesson

Roma non fu fatta in un giorno
(Rome wasn't built in a day)

I thank you.