[From: *Overcoming the language barrier*, 3-6 May 1977, vol.1 (München: Verlag Dokumentation, 1977)]

INTEGRATION OF AUTOMATIC TRANSLATION MODULES IN SPLEEN SOFTWARE. SPECIALLY DEVISED FOR DOCUMENTARY NETWORKS

P. Brossier, A. Mamfredos, C.Milelli, M. Zennaki

Centre National de la Recherche Scientifique
Centre de Documentation Sciences Humaines

Abstract

The Centre de Documentation Sciences Humaines (CDSH) of the Centre National de la Recherche Scientifique was soon aware of the necessity to make its data bases multilingual so as to increase the speed, extent and ease of access to the whole of this information. Parallel to other ongoing trials, an experiment with automatic translation now being conducted relates to the data base of the ENERGY-SAVING documentary network.

This operation has been facilitated by the characteristics of SPLEEN software which, thanks to its modular structure and its functions, allows of automatic translation, both at the level of data input and at that of compilation of an index or of searching.

Two systems have been tested. The first, which is limited to translation of the descriptors, is already operational (for Enqlish and French) and allows of offering several services on the international market: multilingual indexes, retrospective searches and dissemination on profile. The latter, a more ambitious one, is now being implemented. It will allow of the complete translation of the analyses. The difficulties encountered in each of the two cases will be explained, in particular with reference to production of the vocabulary, together with the solutions for these problems.

1. INTRODUCTION

The Centre de Documentation Sciences Humaines (CDSH) has decided to conduct an experiment with automatic translation within the Energy-Saving Network, more specifically on the strength of its data base and the monthly bibliographical review that it publishes.

The foreign language chosen for the first experiment is English.

In a second stage other languages (German, Spanish) will be studied. However, the software is already capable of handling several languages.

Since its creation in 1971 the Energy-Saving Network has drawn up *a* hierarchized thesaurus containing some 3500 descriptors divided into 3 categories:

the concepts (2400)
the products (600)
the geographical countries (350)

2. TRANSLATION OF THE DESCRIPTORS

A Preparation of the descriptors

Before the translation proper, we reviewed our descriptors.
Very useful data-processing tools made the following available to us:

- an alphabetical list of the descriptors (by category) with their frequency of use in the bulletin;
- an alphabetical list of the descriptors (by category) never used;
- a cumulative index allocating the descriptors to the headings of the classification system.

Examination of these lists enabled us to detect certain inaccuracies. We found on the one hand that the thesaurus included terms of exactly the same meaning (e.g. the French expressions for developing countries and underdeveloped countries) and on the other the same word with several meanings ("armement" meaning both the commissioning of a ship and armaments).

Terms were therefore adjusted so that one sole word related to one sole notion.

We also encountered a difficulty due to the impossibility, in the present stage of our study, of taking into account the wealth of characters for input, especially accents.

Indeed, in French the same word assumes a different meaning depending on whether or not it is accented (e.g. arme, armé).

To avoid strict checks, which hamper data collection, we were therefore obliged to chose, according to the frequency of the descriptor, the one which seemed to be used the most frequently, and to find a synonym, an equivalent for the other word so as to be able to express an important notion all the same.

The translation properly speaking was then entrusted to Network specialists to allow each of them to translate the descriptors in his field (gas, electricity, oil etc.). The translation was finally checked by a specialist in English.

The lengthy and detailed procedure that we adopted, i.e. the complete review of the descriptors, proved decisive. The first stage - translation of the hierarchizod thesaurus - made it possible to ascertain that

the translation was correct, after a check on the validity of the link between each of the terms.

B Compilation and utilization of the dictionary

1 - Input of the descriptors and of their translation

- The grammatical details and other data for each French descriptor are as follows:
  . category (concepts, geographical countries, products)
  . gender (masculine or feminine)
  . number (singular and plural, singular alone, plural alone)
  . elision (for words beginning with a vowel)
- The grammatical details of English descriptors are as follows:
  . number
  . the exception of "an" (e.g. an hour)

With the aid of the information supplied by each descriptor, we compiled the dictionary for the translation and the checks on the vocabulary.

The dictionary is composed of two complementary files:

- the dictionary file properly speaking is a direct-access file in which the words and their links are sequentially stored;
- the files of the tables loaded in the memory during processing and containing the addresses of the words in the dictionary file.

To be able to find a word in the dictionary. Fig. 1 shows that it is transformed by hash-coding into two numbers, one, the greater one, serving to distinguish between two words having the same greater number (cf. the following page).

2 - Check programs linked to the descriptor input

These programs post the missing obligatory information:

- **missing plural descriptor if it has not been specified that only the singular existed;**
- **missing gender (for French);**
- **missing code meaning elision when the French descriptor begins with a vowel**
- **missing category.**

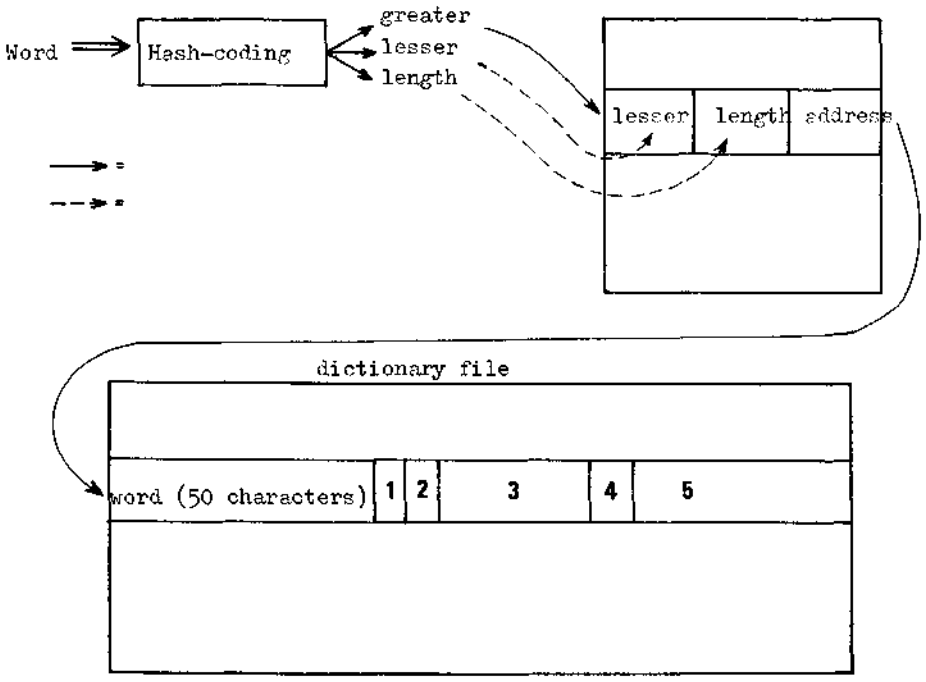
These automatic postings, which occur immediately upon input, make careful rereading of these basic areas unnecessary.

Example 1

052200101 01 1m soffshore

CHECK WHETHER THERE IS ELISION OR NOT (X OR BLANK):x

Figure 1



1 - reserved

2 - word made up of 8 bits with the following meaning:

O-3 nature of the word

4 masculine word if 1

5 feminine word if 1

6 word also used in plural

7 word also used in singular

3 - for the various languages, French, English, Spanish, German, the address is given for the corresponding word in that language (track number and record number)

4 - identical with 3 - for the singular or plural address

5 - identical with 3 - for the address of 7 synonyms.

Example 2

0528012swimming pools? -------▶ ? = request for correction

WHICH CARDS DO YOU WANT TO CHANGE?11

TYPE OF PROCESSING DESIRED : c

PUT A ? UNDER THE FIRST CHARACTER TO BE CHANGED

0528011 SWIMING POOL

?

?ming pool

PUT A ? UNDER THE FIRST CHARACTER TO BE CHANGED

0528011 SWIMMING POOL

3 - Updating the data base

The second stage of our experiment was aimed at translating the descriptors contained in the index to the Energy-Saving Review published by the Network.

A thesaurus can be changed periodically, either to a new descriptors to it or to remove existing ones. That is why, before each updating of the data base, we completed the checking and correction cycle by inserting a new vocabulary checking program, signalling:

- that a descriptor encountered in the text has not been translated. Reference is made to the number of the line in the review where the signalled word appears;
- that a descriptor is not spelt in the same way as in the dictionary;
- that the nature of the word (product, concept etc.) is incorrect.

Example 1

---▶ NON-EXISTENT WORD, LINE: 722 CONFORT CLIMATIQUE

Example 2.

---▶ INCORRECT WRITE, LINE: 398, TYPE: M→MICRO-ORGANISM

the word exists in the dictionary without hyphen

Example 3

---▶INCORRECT WRITE, LINE: 2805, TYPE: N→SOUTES

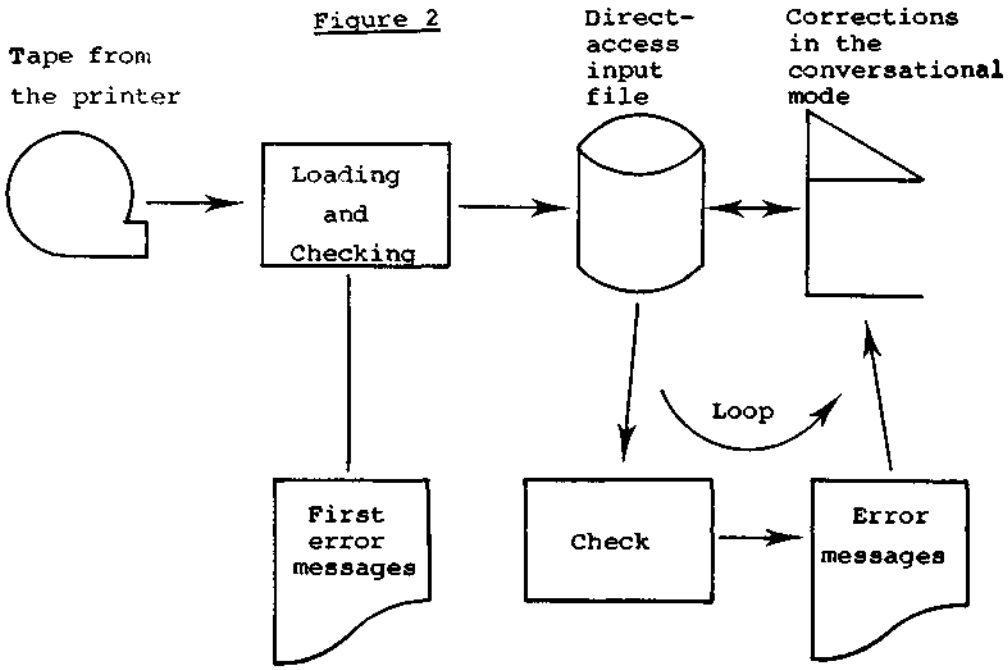
N = incorrect nature.

These check programs make it possible to correct the file and avoid loss of information, for as soon as bilingual indexes are created or the file is interrogated in English, an untranslated or incorrectly written descriptor will not be retrieved.

Moreover, since one error may conceal another one or a correction may have been badly made, it is necessary to recheck the programs. This loop continues until no further error is detected (generally two or three times).

Checking and correction cycle

Figure 2



4 - Special case - synonyms

It is now possible to enter into the dictionary several French synonyms that will be translated by one English word. But if we follow the opposite procedure, i.e. the translation from English into French, the English descriptor will be translated by the first French synonym encountered.

C. Services offered on the international market

1 - Multilingual indexes

The indexes make a quick search possible for foreign users (the descriptors are classified in alphabetical order). They also allow them to make a very fine search, since several indexes are available to them:

- index by concepts
- index by geographical countries
- index by products
- index by records of the classification system

Moreover, it is possible to make up indexes on demand: an index of concepts under their headings etc.

Indeed, the compilation of such indexes presents no difficulties with the SPLEEN system, assuming that it is entirely bounded by parameters. The processing sequence is always the same:

- extraction of the areas from the base
- preparation
- conversion (creation of record so as to be able to sort the areas and sorting)
- packing (so that one needs to edit once only a unit under which several other units come)
- editing.

whatever the index desired, it will be enough to change the instruction tables of the programs for the various stages, except that for the "preparation"

stage that all the special work is done (purification of areas, scatter read, decoding) and in particular translation.

2 - Multilingual magnetic tapes

We can make available to foreign users with a data processing centre a magnetic tape containing a multilingual data base.
The tapes that we can supply are in a format similar to those usually recommended in the various automatic documentation centres. Each posting gives rise to a record.
These records consist of a repertoire giving the address of the areas in the variable part that follows and in which the areas are successively dumped with their length.
For the multilingual tapes the repertoire contains a supplementary key giving the address where the translated descriptors have been dumped.
In this way users can create a tool appropriate to their needs: index, bulletin, retrospective search, selective dissemination of information.

3 - Retrospective search - selective dissemination of information

Two inquiry systems are operational: SPLEEN 2 and SPLEEN 3.

a) SPLEEN 2

This is a system oriented towards selective dissemination of information or selective dissemination on profile.
to use this system starting from questions couched in English only, one has to confine oneself to a search on descriptors, truncated or not to the right or the left. The other possibilities, search on the natural language for the title or other

areas would not be relevant.

b) SPLEEN 3

This is a system oriented towards a retrospective search in the conversational mode that uses inverted files.

Only the areas translated, i.e. the descriptors, types of documents and Readings, are inverted.

The retrospective search on the (translated) descriptors is enough to find the documents that would result from an inquiry on the natural language.

However, this remains tied to the indexing modes.

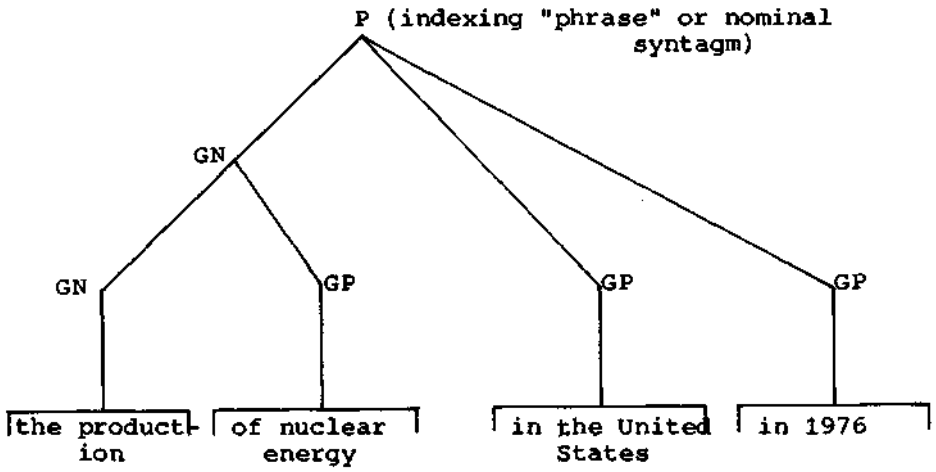
As regards our experiment, we have noted that all the relevant documents appeared, irrespective of the kind of inquiry used (on descriptors, on natural language).

However, in the case of inquiry on natural language, it has happened that a document appeared for the sole reason that the title contained one of the descriptors used in the inquiry, whereas in reality the document did not answer the question.

ELABORATE AUTOMATIC TRANSLATION

The simple structure of the SPLEEN indexing phrases (no conjugated verbs) used by the Network prompted us to study the development of an automatic translation system of a more elaborate nature which would no longer relate to the descriptors only, but to the whole phrase. Indeed, the indexing phrase is only an assembly of words which is self-contained, an explicit rewriting of the original title of a document. It is practically inevitable to encounter in the first phrase (principle phrase) three groups of elements: the subject, the place, the date. The other phrases (secondary phrases) are likewise written in a very simple style and complete the principle phrase.

Example of syntactical analysis of an indexing "phrase"



GN = nominal group generally originating from nominalization on a verbal basis or adjectival basis.
"The production of nuclear energy in the United States in 1976" could be written as "The United States produced nuclear energy in 1976".
The verb "to produce" has been nominalized into "production".

GP = nominal group introduced by a preposition, or prepositional group.

A. Preparation of function words

This second stage proved very complex on account of the nature of the words to be translated: verbs in the infinitive, present and past participles, adjectives, prepositions etc.
Our procedure was identical with that which we explained for the descriptors.
We used an alphabetical list of all the function words and their frequency (about 1000 words) contained in the summaries of our reviews, and we placed them in

their traditional grammatical category.
In the case of some of these words, belonging to several categories at the same time, we had to search for them in their context: the assignment criterion was the frequency of grammatical behaviour of the element.
Parallel to this assigning, we analysed the content of the words. We encountered the same problems as in the study of the descriptors: the same word with several meanings, homonymy owing to the impossibility of taking accents into account.
These problems have been solved as in the case of the descriptors. The English translation was made by the members of the Network, aided by an English linguist.

B. Construction and utilization of the dictionary

1 - Entry of the function words and their translations

The grammatical details for the French word are as follows:

- category (noun, verb, adjective)
- gender, number, elision } according to category

The grammatical details for the English word are as follows:

- category, number, gender } according to category

2 - Checking programs linked with the entry of the function words

A check identical with that for the input of the descriptors has been performed.

3 - Updating the data base

Checks verifying the syntax of the phrases are being studied. At the present stage we may say that the checking cycle at data input will be a little heavier.

However, this overload is minimized by the fact that the checking circuit will remain identical with that of Section 2.B3, Fig. 2. It did not seem necessary to us to install a line-by-line check in the conversational mode at the time of data collection, for the following reasons: the cost is quite considerable, the reliability and the response time are not generally satisfactory and the person doing the data collection, not being trained to the vocabulary and to the various aspects of the data base, would not be able to respond to the messages received.

C. Services offered on the international market

Integration of the phrase translation module for the analysis area with the SPLEEN system is quite possible, This is due to the structure of the data base, which allows of adding new areas without any problem, and also to the structure of the software.
We can store in the base on the one hand the analysis in the original language and on the other the translation of this analysis into a pivot language. Every translation from one language into another would be done via the pivot language, which is an intermediate language. Several magnetic tapes will then be available on which the analyses and the descriptors will be in the language desired. There will thus be tapes in English, in French etc.
The search with the assistance of SPLEEN 2 or SPLEEN 3 (in batch processing or conversational mode) can be done by utilizing all the possibilities; inquiry on descriptors or on the natural language. Editing of the replies will be as required: descriptors only and/or full analysis. As regards SPLEEN 3, the same data base will serve for inquiry and it will be possible to query the base in the language desired and to

edit the summary either in the original language or in the language desired. In this case the response time is slightly longer, the translation being made before display.

4. CONCLUSION

The translation of descriptors has the advantage over the elaborate translation of being operational at present. There is no doubt that the former appeals more particularly to specialized users. For successively aligned descriptors have a meaning for the specialist, who will soon know whether the document signalled in this way is of some interest to him. This reading is less evident to a non-specialist. He will doubtless prefer a more elaborate translated text, even if the syntax is far from Proustian.

When we consider the difficulties that we encounter in making a complete translation of sentences of very simple syntax, we may say that we are very far from the day when the famous aphorism "traduttore, traditore" will be no more than just a saying.