SOFTWARE TOOLS FOR THE ENVIRONMENT OF A COMPUTER AIDED TRANSLATION SYSTEM[1]

Daniel BACHUT - Nelson VERASTEGUI

IFCI
INPG, 46, av. Félix-Viallet
38031 Grenoble Cédex
FRANCE

GETA
Université de Grenoble
38402 Saint-Martin-d'Hères
FRANCE

ABSTRACT

In this paper we will present three systems, ATLAS, THAM and VISULEX, which have been designed and implemented at GETA (Study Group for Machine Translation) in collaboration with IFCI (Institut de Formation et de Conseil en Informatique) as tools operating around the ARIANE-78 system. We will describe in turn the basic characteristics of each system, their possibilities, actual use, and performance.

I - INTRODUCTION

ARIANE-78 is a computer system designed to offer an adequate environment for constructing machine translation programs, for running them, and for (humanly) revising the rough translations produced by the computer. It has been used for a number of applications (Russian and Japanese, English to French and Malay, Portuguese to English) and has been constantly been amended to meet the needs of the users[Ch. BOITET et al., 1982].In this paper, we will present three software tools for this environment which have been requested by the system's users.

II - ATLAS

ATLAS is an aid to the linguist for introducing new words and their associated codes into a coded dictionary of a Computer Aided Translation (CAT) application.

Previously, linguists used indexing manuals when adding new words to dictionaries. These manuals contained indexing charts, sorts of graphs enabling the search for the linguistic code associated with a given lexical unit in a particular linguistic application. The choice of one path in a chart is the result of successive choices made at each node. This may be represented by associating questions to each node and the possible answers to the arcs coming from a node ; the leaves of the tree bear the name of the code and an example.

A language to write the "indexing charts" is provided to the linguist. An ATLAS session begins with an optional compilation phase. Then, the system functions in a conversational way in order to execute commands.

The main functions of ATLAS are the following :

- Editing and updating of indexing charts : compilation of an external form of the chart, and modification of the internal form through interaction with the user, with the possibility of returning a new external form.
- Interpretation of these charts, in order to obtain the linguistic codes and the indexing of dictionaries. A chart is interpreted like a menu, so that the user can traverse the charts answering the questions. He can also view the code found, or any other code, by request, and examine and update the dictionary by writing the code in the correct field of the current record.
- Visualisation of charts in a tree-like form in order to build the indexing manuals.

In the case of interpretation, the screen is handling as a whole by the system : it manages several fields such as the dictionary field, the chart field and the command field.

The system is written in PASCAL, with a small routine in assembler for screen-handling.

Below, we give two examples :

- The first is a piece of tree built by the system based on an indexing chart.
- The second is a screen such as the user sees it in the interpretation phase.

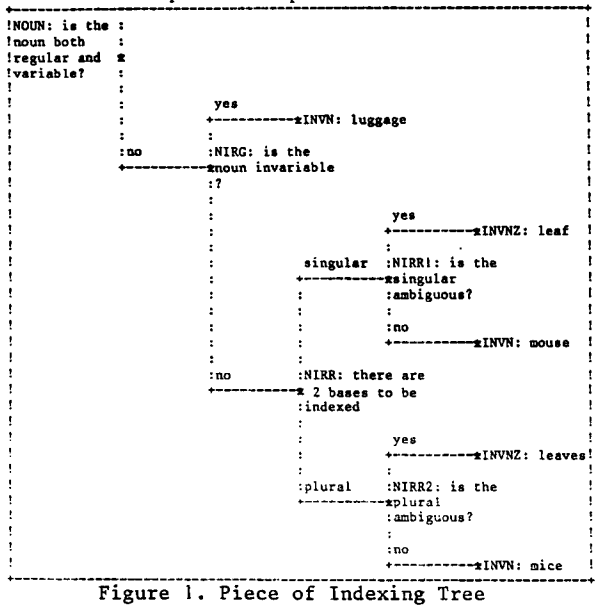

Figure 1. Piece of Indexing Tree

[1] Work supported by ADI contract number 83/175 and by DRET contract number 81/164.

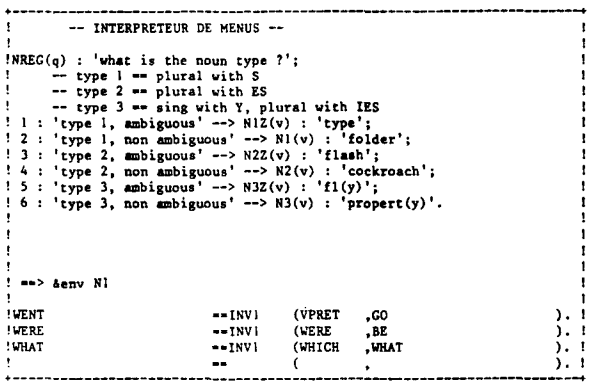

Figure 2. Screen Display during Interpretation Phase.

III - THAM

Computers can help translators in several ways, particularly with Machine Aided Human Translation (MAHT). The translator is provided with a text editing system, as well as an uncoded dictionary which may be directly accessed on the screen. But the translation is always done by the translator.

THAM consists of a set of functions programmed in the macro language associated with a powerful text editor. These functions help the translator and improve his effeciency.

The conventional translation of a text is generally performed in several stages, often by different people : a rough translation followed by one or several revisions : linguistic revision, "postediting", or "technical revision". Hence, the THAM system works with four types of objects : source text (S), translated text (T), revised text (R) and uncoded dictionary (D). In the actual system, each of these objects corresponds to one "file".

The file S contains the original text to be translated, the file T contains the rough translation resulting from a mechanical translation or a first unrevised human translation.

The uncoded dictionary is composed of a sorted list of records following a fairly simple syntax. The access key is a character string followed by the record content, on one or several lines, in a free format. In general, the "content" gives one or several equivalents, but it can also contain definitions, examples, and equivalents in several languages : it is totally free (and uncontrolled).

Finally, the file R is the final translation of the original text realized by the user from the three previous files.

THAM is designed for display terminals. It can simultaneously display one, two, three or four files, in the order desired by the user. The screen is divided into variable horizontal windows. The user can consult the dictionary with an arbitrary character string (which may be extracted from one of the working files), update the dictionary, insert into the revision file a part of another file, make permutations or transpositions of several parts of a file, and receive suggestions for the translation of a word displayed in a window. Moreover, the system can simultaneously use many source, translation, dictionary or revision files.

Basic ideas for THAM come from various sources such as IBM's DTAF system (only used in-house on a limited scale) and [A. MELBY's TWS 1982]. Initial experiments have shown this tool to be quite useful.

IV - VISULEX

VISULEX is a handy and easy-to-use visualisation tool designed to reassemble and clearly distinguish certain information contained in a linguistic application data base. VISULEX is intended to facilitate the comprehension and development of coded dictionaries which may be hindered by two factors : the dispersal of information and the obscurity of the coding. In ARIANE-78, the lexical data base may reside on much more 50 files, for a given pair of language.

This data base is composed of dictionaries, "formats" and "procedures" of the analysis, transfer and synthesis phases (the 3 conventional phases of a CAT system). For any given source lexical unit in this data base, VISULEX searches for all the associated information.

VISULEX offers two levels of detail. At the first level, the information is presented by using only the comments associated with the codes found. At the second level, a parallel listing is produced, with the codes themselves, and their symbolic definition. The first level output can be considered as the kernel of an "uncoded dictionary".

The system provides, on one or several output units, a formated output, with these different visualisation levels.

This system can be considered to have several possible uses :

- as a documentation tool for linguistic applications ;
- as a debugging tool for linguistic applications ;
- as a tool for converting the lexical base into a new form (for instance, loading it into a conventional data base).

It is possible to imagine VISULEX results being used as a pedagogical introduction to a CAT application, seeing that the output form is more comprehensible than the original form.

For the Russian-French application, VISULEX output gives two listings of around 150,000 lines each. This makes it a lot easier to detect indexing errors, at all levels. This is a first step towards improved "lexical knowledge processing".

Finally, we give an example of a VISULEX output. The chosen lexical unit is "CHANGE" in the English-French pedagogical prototype application. The two levels are showed (the left column correspond to the first level, the right column to the second).

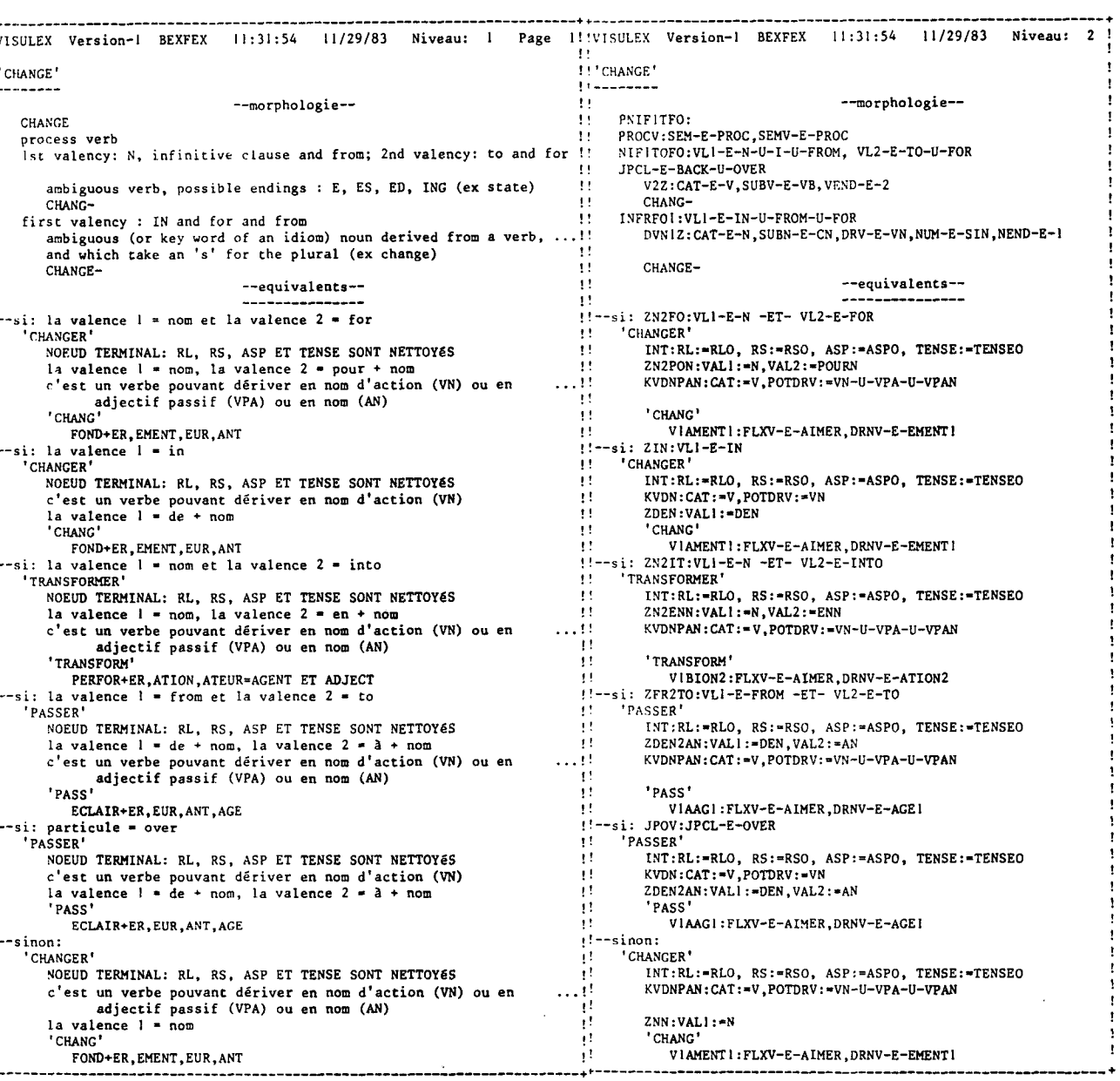

Figure 3. The two levels of VISULEX output

V - CONCLUSION

These software tools have been designed to be easily adaptable to different dialogue languages (multilinguism). The development method used is conventional structured, modular and descending programming. Altogether the design, programming, documentation and complete testing represent around two man/years of work. The size of the total source code is around 15,000 PASCAL lines and 4,500 EXEC2/XEDIT lines, comments included.

The ARIANE-78 system extended by ATLAS, THAM and VISULEX is more comfortable and more homogeneous for the user to work with. This is the first version, and we already have many ideas provided by the users and our own experience for improving these systems.

VI - REFERENCES

BACHUT D.
"ATLAS - Manuel d'Utilisation", Document GETA/ADI, 37 pp., Grenoble, March 1983.

BACHUT D. and VERASTEGUI N.
"VISULEX - Manuel d'exploitation sous CMS", Document GETA/ADI, 29 pp., Grenoble, January 1984.

BOITET Ch., GUILLAUME P. and QUEZEL-AMBRUNAZ M.
"Implementation and conversational environment of ARIANE-78.4, an integrated system for translation and human revision", Proceedings COLING-82, pp. 19-27, Prague, July 1982.

MELBY A.K.
"Multi-level translation aids in a distributed system", Proceedings COLING-82, p. 215-220, Prague, July 1982.

VERASTEGUI N.
"THAM - Manuel d'Utilisation", Document GETA/ADI, 35 pp., Grenoble, May 1983.