[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland;
ed. Kirsten Falkedal (Geneva: ISSCO).]

A Management Tool for Test Corpora

Gerardo Arrarte
Teófilo Redondo
Miguel Sobejano
Isabel Zapata
IBM Madrid Scientific Center

1 Introduction

IBM is engaged in advanced research and development projects on various aspects of Machine Translation between several language pairs, one of which, **LMT (Logic-based Machine Translation),** was started by M. C. McCord in T. J. Watson Research Center in 1985 (McCord 85, 88, 89a, 89b, 89c). The IBM Madrid Scientific Center is taking part in this project through the development of English-Spanish and Spanish-English MT systems.

One of the major challenges in all MT systems is testing and validating the system performance through the use of linguistic corpora (King 90). For the English-Spanish LMT prototype a large and varied set of corpus sentences (around 7500, at this stage) has been selected for the purpose of testing. The sentences were taken from different IBM hardware and software manuals, as well as some other sources such as dictionaries, grammar books and others. Finally, made-up sentences have also been included to cope with specific English-Spanish translation problems.

We have made a special point of developing a tool for building and managing such corpora: the **Corpus Database Manager** (CDBM). CDBM enables NLP researchers to have ready access to the texts in the corpora in a selective way. That means being able to select sentences sharing linguistic features which are implicit in them, but which need to be stated by a linguist. To do this, sentences had to be marked first with a *set of labels* showing each of the features they contain. Both the *selection* (or *retrieving)* and *classification* (or *typification)* processes imply a huge amount of work by qualified experts, which could be optimized by means of the CDBM. This tool takes advantage of Database facilities to store the sentences along with the labels attached to them.

2 Sentence classification

We stated in the previous section that sentences in the corpora have to be marked first with a set of labels. A label taxonomy has been designed so that the expert who will make the classification can go through this taxonomy and easily select the labels that best describe the linguistic phenomena in the sentence.

From a structural point of view, this taxonomy is ranked as a tree in which terminal nodes or *leaves* are the labels themselves and non-terminal nodes represent high-level sets in which labels are grouped.

From a formal point of view, we have made a preliminary classification of linguistic phenomena into a) sentential features; b) verb phrase features and c) noun phrase features.

In relation to sentential features only two important linguistic phenomena have been considered so far: *negation* and *coordination.* We have classified *negation* into *NEGD, NEGP, NEGA, NEGV* for negative determiners, pronouns, adverbs and verbs, respectively. For the *coordination* problem, only coordination between parallel structures has been taken into account, for instance, *ETHOR* for *either ... or, ASAS* for *as ... as*, and so on.

In relation to the noun phrase, we have distinguished syntactic, morphologic, lexical and semantic features. Syntactic features study the use of modifiers and nominal clauses. An exhaustive classification of modifiers has been proposed. It includes articles, possessive and demonstrative determiners, quantifiers, predeterminers, attributive adjectives, prepositional phrases, relative clauses and other related phenomena such as noun clusters, saxon genitive and appositions.

Concerning morphologic features, for the purpose of our research, we have only taken into account the use of personal, demonstrative and possessive pronouns. Lexical features refer to the occurrence of particular lexical units. In this case, the pronoun *it* raises enough problems so as to deserve special consideration. With respect to semantic features, a broad study would be necessary to achieve a more complete classification, but in the current LMT application we do not have to face so deeply with semantics.

For the verb phrase we have also distinguished syntactic, lexical and semantic features. The syntactic classification is far more exhaustive than for the noun phrase. What is studied under syntactic features is the use of imperative, infinitive and gerund as non-finite forms of the verb; subjunctive mode; perfect and progressive tenses, passive voice, elliptic passive[1], *be going*

[1] Sometimes in English a passive meaning is expressed by means of a mere past participle. In these cases, different syntactic structures have to be generated in Spanish to translate this meaning. For instance, a relative clause or, like in the example below, a full passive: *Users can access other user's objects only if authorized.*

Los usuarios pueden acceder a los objetos de otro usuario slo si son autorizados.

to and *be about to* periphrases and modal verbs to describe verbal aspect.

Another important syntactic feature of verbs is their argument structure. We have distinguished the occurrence of direct object, indirect object, predicative adjectives, agentive and adverbials. The latter can be of three types: clauses, prepositional phrases and simple adverbs.

Lexical phenomena include the use of verbs *be, have, there be* and other special verbs.[2] As for semantic features, we intend to study here semantic types of verbs and the semantic requirements of their arguments (subject, direct object, etc). As well as for noun phrase semantic classification, further work has to be done on this verb and argument typology.

3 Functional Description

3.1 Overview

The CDBM tool has been developed using the "C" and "SQL" languages under the OS/2 operating system. It takes advantage of OS/2, Presentation Manager and Database Manager functions and capabilities.[3]

Essentially, CDBM is an application program that provides various facilities to easily create and maintain a specific kind of database. The structure of a CDBM database can be seen as a table. Each row of this table holds a sentence along with a set of *labels* representing its linguistic features.

From the user's point of view, there is no need to know anything about the internal structure of a CDBM database. The program provides all the basic operations needed to create and maintain such a database in a simple way:

- **Create** and **delete** a database with CDBM structure.
- **Select** one active CDBM database from within the existing ones.
- **Retrieve** from the active CDBM database sentences that share some specific labels and edit them.
- **Select, classify** and **store** sentences from an edited text file.

[2] We call *special verbs* those verbs with special argument structures. For example, the raising *movement* with the verb *want* in

I want him to go.

A farther subclassification will be done of this special verb group including labels for each verb considered.

[3] These products are under IBM Corporation Copyright. OS/2 is an Operating System; Presentation Manager is a window management user interface; Database Manager is an SQL Database interface.

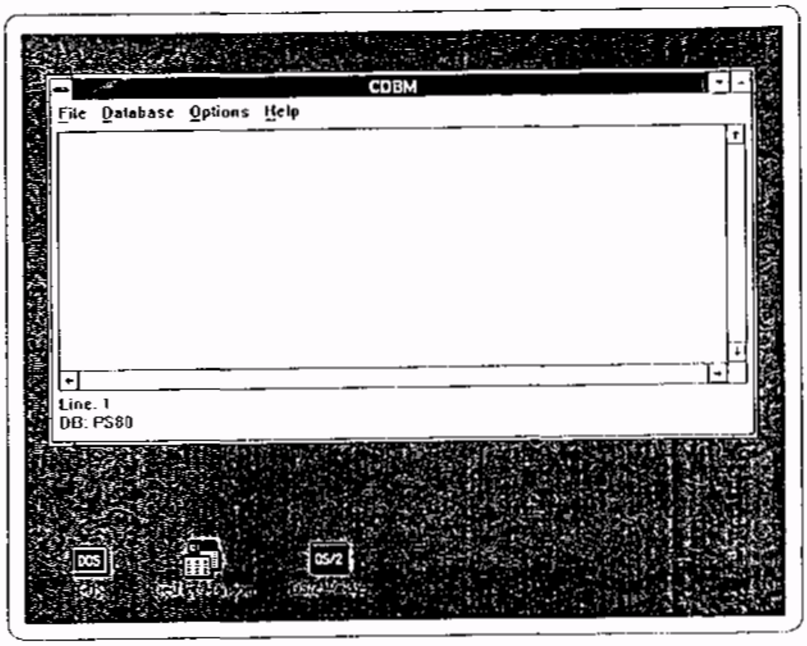
CDBM has been designed to be as easy to use as possible. The overall program operation makes extensive use of windows, menus and dialogs, and all options can be selected with the mouse as well as with the keyboard. From the beginning to the end of the program, dialog-driven menus lead the user through a set of choices until each operation is completed.

4 Using CDBM

The following sections will show the general operation of CDBM, with its main features.

4.1 Starting CDBM

When started, the CDBM Tool displays its main window. The files from which to select sentences and the files where the retrieved sentences are to be saved will be browsed here. The name of the active database is shown at the bottom of the window.



Line. 1
DB: PS80

Figure l.CDBM Main Window

This main window contains an action bar, showing the names of some pull-down menus, from which several choices can be made. See fig.l for an illustration.

The File pull-down menu allows the following choices:

- Classify Sentences
- Retrieve Sentences
- Exit the program

The Database pull-down menu allows the following choices:

- Create a Database
- Select the active Database
- Delete a Database

4.2 Database operations

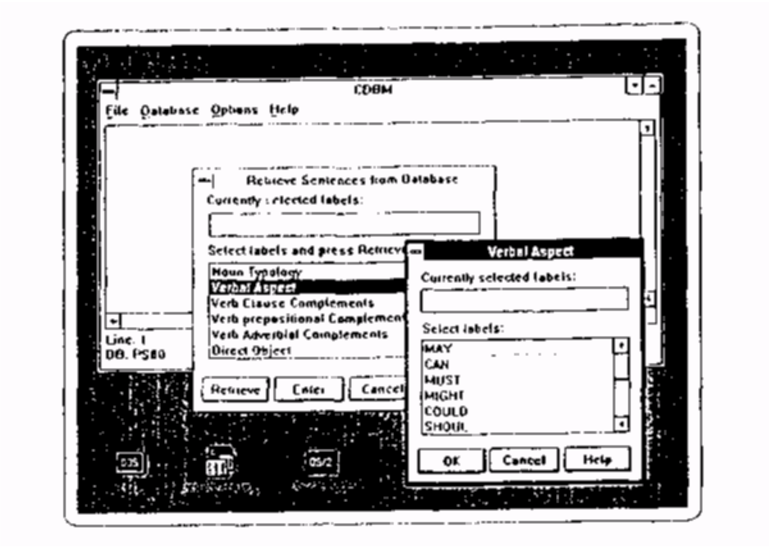
To allow the organization of the data, different sets of sentences belonging to different corpora can be stored in different databases. Therefore, the first thing that should be done after starting CDBM is to select an active database. If no databases exist yet, a new database can be created. A database can also be deleted if it is no longer needed.

When the **Select Database** or the **Delete Database** options are chosen, a list with all the existing databases is shown. The selection may be done by pointing to the desired database and then pressing <ENTER> or doubleclicking the mouse.

4.3 Retrieving sentences

Once a database has been selected, you can retrieve menu sentences from it by selecting the **Retrieve** choice from the **File** pull-down. A window appears with a list of groups of linguistic features. You have to decide whether you want all the sentences to be retrieved or only those which share some selected attributes. To select an attribute you first select a group from the list, and then the desired label from the corresponding label list.

The group and label lists are shown in fig.2. Finally, after pressing the **Retrieve** button, the sentences found are shown in the main window, from where you can save or discard them. If you did not select any labels, all the sentences in the database would be retrieved.

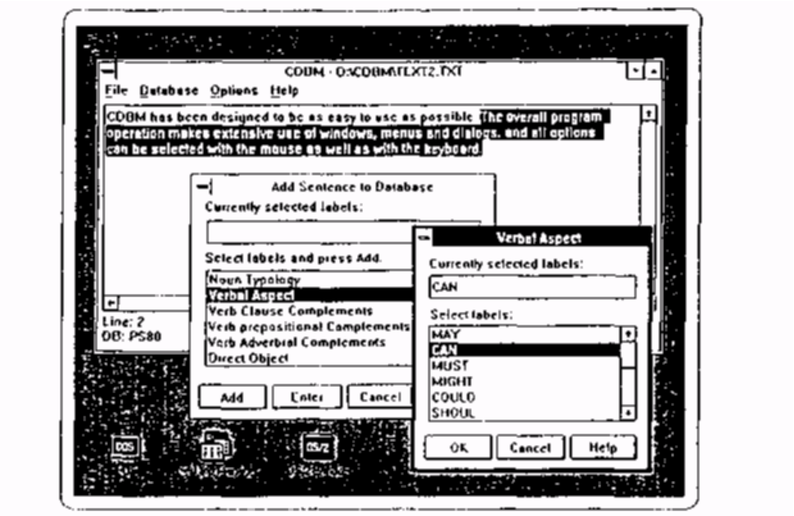

CDBM
File Database Options Help
Retrieve Sentences from Database
Currently selected labels:
Select labels and press Retrieve
Noun Typology
Verbal Aspect
Verb Clause Complements
Verb prepositional Complements
Verb Adverbial Complements
Direct Object
Retrieve
Enter
Cancel
Verbal Aspect
Currently selected labels:
Select labels:
MAY
CAN
MUST
MIGHT
COULD
SHOUL
OK
Cancel
Help
DOS
OS/2

Line: 1
DB: PS80

Figure 2.Retrieving Sentences

4.4 Classifying Sentences

To classify a sentence you select the Classify option from the File pull-down menu. Then you have to select the name of a file containing sentences to be classified. After that, the file is shown in the main window. When the mouse is doubleclicked, a sentence is highlighted. A list with groups of linguistic features appears. If the highlighted sentence already exists in the database, the stored labels are shown and you can update them. If the sentence does not exist in the database, you have to select the desired labels from the different groups and when you are done, press the Add button to store it.



CDBM has been designed to be as easy to use as possible. The overall program operation makes extensive use of windows, menus and dialogs. and all options can be selected with the mouse as well as with the keyboard.

Line: 2
DB: PS80

CAN

Figure 3.Classifying Sentences

Figure 3 shows an edited text file with a highlighted sentence and both the group and label list windows.

4.5 System Architecture

The CDBM program has been developed using the OS/2 1.3 EE operating system. It has been fully programmed using the C programming language, and internally uses the SQL language to communicate with the OS/2 Database Manager.

The CDBM program is divided into modules:

- Main module
  Depending on the user input, the main module passes control to the other modules of the program.

- Browser
  The main window includes a simple browser used to show both the retrieved files and the files from which to classify sentences. It has been developed using the PM API (Presentation Manager Application Programming Interface).

- Database modules
  For each of the main database operations: create and delete a database and store and retrieve data, a module exist. Each of them have to pass a number of SQL commands to DBM (Database Manager) using function calls. A dialog driven user interface is used to make the SQL syntax transparent to the user.

5 References

(CEE 83) Commission of the European Communities “Better Translation for Better Communication”, Pergamon, Oxford.

(Guida 84) G. Guida, G. Mauri “A Formal Basis for Performance Evaluation on Natural Language Understanding Systems”. *Computational Linguistics,* 10, 15-30.

(King 90) M. King, K. Falkedal “Using Test Suites in Evaluation of Machine Translation Systems” in *Proceedings of the 13th International Conference on Computational Linguistics,* Helsinki, 1990.

(Lehrberger 88) J. Lehrberger, L. Bourbeau “Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation”, John Benjamin, 1988.

(McCord 85) Michael C. McCord "LMT: A Prolog-Based Machine Translation System" in *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* S. Nirenburg (Ed.), Colgate Univ., 1985.

(McCord 88) Michael C. McCord "A Multi-Target Machine Translation System" in *Proceedings of the International Conference on Fifth Generation Computer Systems,* 1988, Institute for New Generation Computer Technology, Tokyo, 1988.

(McCord 89a) Michael C. McCord "Design of LMT: A Prolog-Based Machine Translation System". *Computational Linguistics,* 15, pp. 33-52, 1989.

(McCord 89b) Michael C. McCord "LMT" in *Proceedings of the MT Summit II*, Munich, 1989.

(McCord 89c) Michael C. McCord "A New Version of the Machine Translation System LMT". *Literary and Linguistic Computing,* vol. 15, pp. 218-229, 1989.

(Slocum 85) J. Slocum, W. Bennett "An Evaluation of METAL: the LRC Machine Translation System" in *Proceedings of Second Conference of the European Chapter of ACL,* 1985.