CONTENTS

In this issue 38
Books from Sesame 39

Articles

Japanese Text Processing and
Electronic Mail on the IBM PC and Macintosh
Troy Dillard and Ken Lunde 40

ADMYTE: The Digital Archive
of Spanish Manuscript and Texts
Francisco A. Marcos Marín,
Aurora Martín de Santa Olalla,
Charles B. Faulhaber and
Ángel Gómez Moreno 50

A Classification of Scripts
John Clews 63

Software Overview

The Multilingual PC Directory, update 2
Ian Tresman 76

Book Review

Characters and Computers
Editors: Victor Mair, Yongquan Liu
Reviewer: A.W. Rowe 80

From East to West: Automation News
84

Sesame Bulletin is abstracted in:
Linguistics and Behaviour Abstracts (LLBA) and
Library and Information Abstracts (LlSA)
IN THIS ISSUE

There is an increasing trend to provide increased access to the heritage of a particular culture to be available to scholars and others, formerly on-line, or on a series of computer discs, and increasingly on CD-ROM. This trend will continue to be dealt with by SESAME Bulletin. Interestingly, it has mainly been older non-roman script materials which have made the greatest impact. One of the earliest ventures was the Thesaurus Linguae Graecae founded almost two decades ago, largely by Theodore Brunner, Professor of Classics at the University of California at Irvine. Its six million words contain the entire corpus of classical Greek literature from Homer to the 6th century A.D. A more recent development has been plans for a Sanskrit database, with possibly ten times as much literature available as in the Thesaurus Linguae Graecae. The Asian Classics Input Project has already been making Tibetan and Sanskrit works available for some years, both in transcription and also enabling them to be shown in Tibetan script. The Research Libraries Group has also been involved in collaborating with East Asian institutions in recording details of the heritage of China, Japan and Korea, and the development of EACC (the East Asian Character Code) was developed as part of this effort.

Perhaps surprisingly, Latin language corpora have been slower to appear, particularly in Europe, although this trend is now reversing. Classical and ecclesiastical collections like Jacques-Paul Migne's Patrologia cursus completus are now appearing, and there is now a very large number of translations of the Bible, and other sacred texts like the Qur'an available. The heritage of modern European languages is also appearing on compact discs. The Domesday project in the UK provided an audiovisual record of past and present culture in Britain, albeit limited to older 12" optical disc technology which predated CD-ROMs. Oxford University Press together with Longman and Chambers, Oxford and Cambridge Universities and Lancaster University's Unit for Computer Research on the English Language, now aim to produce a machine-readable corpus of 100 million words of contemporary spoken and written British English.

In Spain, with 1992 marking the quincentenary of Christopher Columbus reaching America, the Sociedad Estatal para la Ejecución de Programas del Quinto Centenario has been collaborating with other institutions in Spain and America to develop ADMYTE: the Digital Archive of Spanish Manuscripts and Texts (see pages 50–61) in order to provide access to the entire corpus of medieval Spanish literature. The second article in this issue provides a very useful overview of the sort of collaboration and planning required to develop such a venture.

The other major focus from this issue onwards is East Asian language automation. The opening article provides information that is very difficult to get hold of outside Japan, namely Japanese Text Processing and Electronic Mail on the IBM PC and Macintosh by Troy Dillard and Ken Lunde (pages 40–48).

A classification of scripts (pages 63–75) by John Clews provides an aid to all those working with languages who need to identify which scripts are used for which languages. Ian Tresman's update to his Multilingual PC directory are continued on pages 76–78 and From East to West: automation news is on page 84. A. W. Rowe's review of Characters and computers (pages 80–83), based on the conference on East Asian Information Processing held in Pennsylvania in October 1990 continues the East Asian emphasis: copies of this book will also be available from SESAME Computer Projects (see next page).
ADMYTE:
The Digital Archive of Spanish Manuscripts and Texts

Francisco A. Marcos Marín, Aurora Martín de Santa Olalla,
Charles B. Faulhaber and Ángel Gómez Moreno
Sociedad Estatal para la Ejecución de Programas del Quinto Centenario
Aravaca, 22 bis, 28040 MADRID, SPAIN

The preservation of their cultural heritage has become a major preoccupation of nations around the world. Books, especially older manuscripts and early printed texts, are at risk because of their rarity and fragility. Spain throughout her history has lost magnificent libraries, such as that of Ferdinand Columbus (the son of Christopher), which today would be the best library in the world for the study of the medieval and early modern literature of the Romance languages had it not been gradually destroyed by neglect. But preservation is not enough. We must also make it available to the world at large, to scholarly readers in particular but also to the general cultured public, that which has been preserved. Unfortunately, these two goals are incompatible: to make books available for use is to ensure their inevitable destruction. It is the dilemma faced by all librarians. Modern technology, however, has given us the means to solve both of these problems simultaneously, to preserve the past and yet to make it available. We intend the Digital Archive to be the first of what we hope will be a series of similar ventures.

With this project we can provide access to the entire corpus of medieval Spanish literature, understanding the term "literature" in the widest possible sense, and, in addition, we can provide along with this corpus a set of research tools, already existing or in advanced stage of development, with which scholars can carry out their studies with total safety to the physical works themselves. Because of that, the Sociedad Estatal para la Ejecución de Programas del Quinto Centenario and the company MICRONET S.A., are carrying out the project ADMYTE. MICRONET has a wide and recognized experience in the field of information storage and retrieval in CD-ROM. Other language industry companies and academic institutions are also collaborating, in particular the Autónoma and Complutense Universities of Madrid in Spain, California in Berkeley (USA), Madison in Wisconsin (USA) and Toronto in Canada, together with the major libraries in the Hispanic World, such as the Biblioteca Nacional in Madrid and in Buenos Aires, and public and private firms.

The participation of those institutions involved from the launch of the project is based on their previous experience in developing products such as ADMYTE and their capacity of managing the tools needed in its development.
- The Sociedad Estatal para la Ejecución de Programas del Quinto Centenario participates through its area of language industry in the coordination of activities aiming at the development of language engineering in Spain.
- MICRONET, S.A. is a private firm specialized in storage and retrieval information in CD-ROM. It has created the CD-ROM edition of data base of the CSIC, Consejo Superior de Investigaciones Científicas and that of the Spanish ISBN. For the information retrieval process it has developed a well-known data base called KNOSYS. MICRONET contributes to ADMYTE with a version of CLARITY-CD, a specific program to retrieve text and image information from CD-ROM. An experienced team of computer technicians will work together with the linguists provided by the Sociedad Estatal and the Universities by installing hardware and developing software for image processing and text recognition.
- The Biblioteca Nacional de España contains not only the best collection in the world of Spanish manuscripts and incunabula, but also has an experienced team of library
professionals whose participation is most needed in selecting the incunabula and old printed books to be included and in revising the bibliographic information.

- The Universidad Autónoma de Madrid houses the Madrid team for the EEC's EUROTRA project in machine translation, the IBM Scientific Centre, and the Institute of knowledge Engineering, also in collaboration with IBM. It is also the centre for the UNITE project for the automatization of textual criticism, whose software will become an integral part of ADMYTE.

- The University of California at Berkeley, the home of BOOST, is one of the leading institutions in the United States for the study of Medieval Spanish; it is also a world-renowned centre for information and computer systems technology.

- The Universidad Complutense de Madrid is participating actively in both BOOST and DOSL, chiefly in the process of collecting and verifying data for both projects.

- The Hispanic Seminary of Medieval Studies, a non-profit institution based on the University of Wisconsin, Madison, is so far the largest disseminator in the world of medieval Spanish texts in non-traditional (microfilm and electronic) format as well as a very active publisher of printed editions. Virtually all of these texts have been transcribed as by-products of its major research project, the Dictionary of the Old Spanish Language, which will also be integrated into the Digital Archive. The Seminary has had a long experience (since 1972) in the use of computers in text processing, lexicography, and typesetting (both for microfiche and paper).

- The University of Toronto is the site of the Centre for Computing in the Humanities, which is one of the most prestigious in the world. One of its product is TACT, a tool for text retrieval and creation of a data base system. This program is included in ADMYTE with specific modifications for this collection.

The main result of ADMYTE is a collection of CD-ROM laser disks, making the most of modern technology. It can be used by scholars who have access to an MS-DOS personal computer with a VGA monitor and a laser disk reader. ADMYTE is divided into two volumes of differing size. The first comprises only one disk which is aimed at researchers with more sophisticated needs. The first volume (disk 0, or tool disk) contains the following:

- BOOST, the Bibliography of Old Spanish Texts (BETA in the Spanish abbreviation) which is a bibliographical data base with more than three hundred fields and eleven interactive tables. This was produced in collaboration with the University of California in Berkeley and the Complutense in Madrid, with the support of IBM USA.

- FORM-LEX, the Dictionary or lemmata and forms, is part of the Dictionary of Old Spanish and was largely developed by the Hispanic Seminary of Medieval Studies of the University of Madison in Wisconsin.

- TEXTS-MAD is a collection of Medieval Texts, transcribed by the Hispanic Seminary of Medieval Studies in collaboration with the University of Madison in Wisconsin. BOOST, FORM-LEX and TEXTS-MAD have also received support from the National Endowment for the Humanities.

- TACT, a text retrieval information program provides a system allowing the creation of its own textual data base developed by the Centre for Computing in the Humanities of the University of Toronto, with the support of IBM Canada.

- UNITE is a group of programs used to construct machine-readable critical editions, developed by the Universidad Autónoma de Madrid, with the support of IBM Spain, IBM Deutschland, EUROTRA-Spain and, the Alexander von Humboldt Stiftung.

The second volume, (containing disks from 1 to n) is aimed at a wider sector of people including, naturally, the scholars themselves. This comprises:

- Transcribed texts in ASCII code, with mark-up or standardized tags.

- Images (facsimiles of texts in black and white and miniatures reproduced in colour).

- CLARITY-CD, a MICRONET retrieval program for texts and images of high resolution.
We are determined to carry out this project not only because of the urgency of preserving Spain's cultural heritage, but also because of the deep change which has occurred in the working habits of humanities scholars. The computer is now widespread among humanists and has become indispensable for certain functions as the amount of information available to researchers increases exponentially. A corresponding development of computerized research tools has become an absolute necessity. For this reason, the scholars who have organized the Archive are conscious of the need of joining together our respective research projects in a synthesis of our experience and knowledge in order to create a vademecum for the 21st century. The programs, techniques, and technologies we have developed and will develop can be applied not only to the language and literature of the Medieval Kingdom of Castile and Leon and its modern descendants but also to languages from any corner of the world. The is also true for the techniques developed for the digitization of MSS or the automatic transcription of incunabula: they will be placed at the disposition of scholars of any language, country, or historical period.

Each of the scholars involved in the project approaches it with a particular set of skills and from a particular point of view; but all share the following aims:

1. Developing research tools as an expert system for the processing of texts and the retrieval of information contained in them.
2. Presenting a model focused in principle on Medieval and Early Modern Spanish but easily extensible to other periods and languages.
3. Recovering Spain's cultural heritage through locating, cataloguing, preserving and studying fundamental aspects of Spanish history: medieval and early modern texts and the material witnesses in which they are found, manuscripts and incunabula.

The Archive project, in collaboration with the National Library of Spain, will also carry out, on a more technical level, a plan for the conservation of manuscripts and incunabula editions by means of their digitization from microfilm and subsequent image processing, followed by the use of optical character recognition programs for the automatized transcription of early printed texts. This builds on a solid foundation of previous work by the collaborating scholars, will also perfect existing research tools, for example amplifying UNITÉ to make it suitable for the treatment of any kind of text in prose and verse, in any language, and stimulating the collection of materials to make of BOOST an exhaustive inventory with a versatile, sophisticated, and easy-to-use interface for non-technical users.

Our overall aim is (a) to make a CD-ROM collection incorporating the whole of BOOST and its system, the collection of Medieval Spanish texts transcribed by the University of Wisconsin (ten thousand pages, many of them unpublished, including Alfonso X el Sabio works, the XV century Castilian collection of lyrical poems or the medieval medicine treatises, among others), and (b) to provide sophisticated computer systems to access and process these texts. These systems include TACT, a system of text retrieval and analysis information, the text-editing program UNITÉ (used for the first time by Francisco Marcos Marín in his edition of El Libro de Alexandre) and MICRONET's CLARITY-CD system to retrieve the information contained in the laser disk. Each disk will be full of digitized facsimiles (for this purpose the collaboration of La Biblioteca Nacional, whose incunabula will be published in the collection, is essential). The whole would be a corpus which can not be found in any library of the world. This would be the contribution of the institutions involved in ADMYTE to the Quincentennial of the discovery of America.

The major benefits of this project are that we will be able to bring the study of the Spanish culture to any place, far away from the big libraries. Equally, the use of
facsimile will contribute to the preservation of the original texts. Spain will become a precursor in the use of a new technology that will have to be adopted soon. Finally, image processing techniques originally developed to improve photographs obtained from space satellites will allow scholars to restore images from manuscripts electronically, so that we will be able to recover an enormous collection of texts hitherto considered illegible.

Specific sections of ADMYTE

This section explains briefly the previous researches about the different aspects involved in the Digital Archive, especially those carried out by the participants, and how we plan to go on with the task. As can be seen, in the majority of the cases, the projects of most importance so far correspond precisely to the scholars assembled in the present effort.

BOOST: Bibliography of Old Spanish Texts

Cataloguing and preserving the testimony of its past is one of the most urgent tasks which a country with a rich cultural heritage must undertake; this is also one of the main tasks for the Spanish Middle Ages specialists. Our experience in the manuscripts and incunabula field has made us insist plenty of times on the necessity of this task: we do not have notice even of the existence of many books kept in public and private libraries. Among these libraries, the most important ones such as the Biblioteca Nacional, Palacio Real, Academia Española, Colombina de Sevilla or Universitaria de Salamanca have not made complete, exhaustive catalogues): a lot of Spanish manuscripts suffer a strong damage and are threatened by several dangers (problems of the corrosive inks are specially serious, although other elements should not be forgotten: the simple flow of time, fires, flooding, mutilations, applications of reactive by ancient readers, moth and fungus); finally we must point out that, in spite of the control of the legislation in defence of national heritage, there are many codex and printed books that slip out the control of the State because of the ignorance of their existence (we have a lot of scandalous examples in Spain and abroad).

BOOST (BETA in the Spanish abbreviation) intends to solve a great part of these problems. BOOST is a catalogue of the primary sources - manuscripts and printed— for the study of medieval Spanish culture, a detailed inventory and identification of the texts they contain, and a biographical data base of all the figures involved in the creation, dissemination, and the use of those sources and texts. By "primary sources" we mean all non-notarial documents written in Spanish or in any of its historical dialects. BOOST began as a spin-off from the Dictionary of the Old Spanish Language in 1974 but soon proved its value as a union catalogue of medieval Spanish manuscripts and printed books. There are a lot of scholars who have collaborated in BOOST from its first edition (966 entries in 1975) to the third one (3,378 entries in 1984). At this moment, Charles Faulhaber and Angel Gómez Moreno have incorporated more than 2,000 new entries to the data base and we think BOOST will have 10,000 references at the moment of its publication (planned for 1992).

As with other projects prepared and developed in Wisconsin University, the BOOST interest also lies on the use and development of new technologies in the studia humanitatis field (some of the members from the international team collaborating in BOOST are consultants in projects developed by computer assistance). One of the major changes of BOOST is the program we use for the data base management. Two years ago, thanks to help from IBM for a demonstration to King Don Juan Carlos and
Queen Doña Sofía, we were able to use the Advanced Revelation relational data base management system on an IBM PS/2. This new system, in ADMYTE, will allow us to prepare not only the fourth printed edition but also the first magnetic one. Whereas the access to the data of the printed edition has to be limited, without choice, to the normal index (onomastic and subjects) the access to the computerized version offers multiple possibilities to the user. The new system is able to search information through more than 450 combined elements in an interactive way with eleven tables presented in an integrated initial menu.

For ADMYTE, BOOST means the enlargement of old texts cataloguing through the following tasks:

1. Reviewing the collections of the Biblioteca Colombina and Capitular de Sevilla; the Biblioteca Universitaria de Salamanca; the Biblioteca Nacional de Lisbon; the Torre do Tombo in Lisbon; the Biblioteca de Ajuda; the Biblioteca de Mafra; the Biblioteca Pública de Porto; the Biblioteca Pública de Coimbra and the Biblioteca Universitaria de Coimbra.

2. Examining catalogues and printed inventories of codices and incunabula, as well as secondary printed sources for the obtention of news about authors and works.

3. Updating the data base to incorporate the results of these investigations.

**DOSL: Dictionary of the Old Spanish Language and Alfonso X**

In Madison, in 1931, the famous scholar Antonio Solalinde founded the Seminary of Medieval Spanish Studies, a centre of attraction of all hispanists of the world. The reason of that is the great amount of information the Seminary has because of the projects they develop; among them, the most ambitious one is the dictionary of Medieval Spanish. This is, summing up, the history of this dictionary, the *Dictionary of the Old Spanish Language*, also called DOSL (DEA in Spanish). After the publication of the first part of Alfonso X General Estoria in 1930, Solalinde began the edition of its second section. They count with the collaboration of two young medievalists: Lloyd Kasten and Victor Oeslischläger. The idea of creating a dictionary of medieval terms appeared in this moment as a task joined to the edition of Alfonso X work, a relationship which is still going on. Antonio Solalinde died in 1937 and Lloyd Kasten became the director of the Seminary. In the following decade the group of the Seminary continue with the task of compiling a dictionary of medieval Spanish and they published as an example the *Tentative Dictionary of Medieval Spanish* (1946). During almost forty years, the group of professor Kasten after dedicating thousands of hours to the work, collected several millions of quotations of Alfonso X which are kept in 564 big drawers. It was also created a dictionary of terms extracted from the printed editions, which, in spite of its great utility, is still unpublished, although any scholar can access to it.

In 1972, John Nitti, one of Kasten's pupils, started his work in the seminary, and in 1975 he became co-director of it. This scholar was already known because of his involvement on the so called new Cybernetics techniques. The union of this two scholars changed the method of compiling DOSL: from 1972 to 1974 several specific programs were created for the Dictionary and the transcription of the microfilms of the manuscripts from the Royal Scriptorium of Alfonso X began. The use of primary sources, more reliable than the printed editions, is the greatest innovation of the Dictionary of the University of Wisconsin.

Alfonso X is the nucleus of DOSL, but this corpus is not the only one to be included, its editors were determined to include 250 manuscripts and incunabula of different contents, written between the XII and XV centuries. Because of that, one of the
basic and first tasks was to locate the primary sources, a difficult one because there
was not a general catalogue with pieces of information about codex and preserved
printed sheets, their localization and their content. The Bibliography of Old Spanish
Texts was born in 1975 with the purpose of carrying out this task, although its great
utility to the knowledge of medieval Castilian texts made it be more than a DOSL tool:
Currently BOOST is the first book where any scholar who needs to study the Edad
Media through its primary sources can turn to.

The Seminary has also had another important influence, the creation of the standard
transcription: the texts are transcribed according to the book prepared by David
Mackenzie, A Manual of Manuscripts Transcription for the Dictionary of the Old Spanish
Language [1986: 4.ed.], and the different entries are created according to the patron
of Victoria Burrus, A Procedural Manual for Entry Establishment in the Dictionary of
the Old Spanish Language [1986:3.ed.].

UNITE

UNITE is a program suite that compares different versions (text tokens) of a given
text type in order to obtain a unified version. It is not intended as a definitive
solution to all of the problems of textual criticism (texts of variable length, existence
of different versions, etc.), but rather as a tool to free the editor from drudgery in
order to allow him or her to concentrate on those aspects of the process which
require human judgment. The system was designed for strophic texts in verse, but it
is currently being reworked for non-strophic texts as well as for prose. Written in
PASCAL, it was originally implemented for IBM 43xx mainframes and above. It has
been ported into UNIX. The Archive's version has been re-written for MS-DOS in C.

The first of UNITE's characteristics is that the editor need add no mark, or label to
the text being transcribed; no such pre-editing is required/textual transcriptions upon
which it is based require no mark-up and in fact can work on straight ASCII texts
with no mark-up at all. The second is that it presents a wide range of possibilities
which allow the user to tailor the suite to his or her own needs through an
interactive menu system.

The standard package is designed for the combined collation of up to ten versions.
Its scope can be enlarged if more than ten texts are to be collated simultaneously.
The logical unit of collation is the stanza, understood as a set of lines of text
separated by blank lines from other stanzas (a poem may be composed by one stanza
of n lines). This type of unit permits the scholar to detect and solve problems of
misplaced lines inside a stanza. It is not necessary for the versions to have the
same number of stanzas neither for them to be arranged in their numerical order,
because various utilities are included which can format and put the texts in the right
order. It is not necessary for the stanzas to have the same number of lines, and
unlike the previous versions, this number has no more limitations than those derived
from the disk capacity.

UNITE is implemented in C, and is presented to the user in a integrated environment by
means of an integrated menu system with a main menu and a number of secondary
menus. By modifying the parameter values, the user is able to control the level of
automation of the unified version. The values contained in the different parameter
files are the default values when the option for modifying them interactively from the
screen is not used. It is possible to view the contents of any file included in the
package by using the computer-editor. The scholar can modify any aspect of the text
files as needed. Extracting variants in a file allows to have all of the variants which
have been produced when making the unification, to know what has been unified, step by step, and to be able to have this data when making the critical program.

One option enables searches of word sequences to obtain information about lexical or syntagmatical forms which could be interesting, inside its verse. In each case the search is done within a file, and the results are stored in a common file to which the data of the subsequent searches are joined, if is that our purpose. Concordances are obtained with an indicative of where the searched words sequences are found, the number of the line, and optative suppression of the differences between capitals or small letters and blank spaces among words which can be joined or separated (such is the case for clitics and other forms). The number of options and combinations is high. Therefore, it is worthwhile to emphasize the choice between getting the concordance in the text or only the numerical indication of the line in which it is found, a faster and more interesting procedure for statistical counts. In ADMYTE, the associated program, par excellence, is TACT.

Another option in UNITE permits the user to format automatically and put the stanzas in order. The format, passed to the unification program either from the screen or by modifying the parameter file, must be the same for every version to be collated. It is important to note that, although a number of extra lines at the beginning of the file is required, those are not modified by means of the format, and only indicate the beginning of the text that is being formatted. Formatting the text according to the values mentioned above also establishes the separation between two stanzas by only a blank line.

The program will search for text files in any directory, provided that the file name includes the full path name to access that file. This ability can be extended to all those utilities of UNITE which are given file names. The original spellings of the texts are modified towards more usual or neutral graphic forms at this stage of the process. This is intended to eliminate as much as possible mere orthographical variation differences in order to obtain the maximum number of identical words during the process of collation.

Spelling unification is divided into five phases. A number of modifications to the original spellings are applied in each one. These phases are not executed one after the other; rather, a process of collation is inserted after each one so that the original spelling of the texts is preserved as much as possible. Thus when a word becomes standardized in one of the modifications, it is not altered by the following ones. There is also a phase called inactive where the original spelling cannot be modified. Its execution at the beginning of the process permits the collation of the texts with the original spelling cannot be modified. Its execution at the beginning of the process permits the collation of the texts with their original spellings, and prohibits subsequent modifications of words so collated. The user is able to determine what phases will be executed and in which order, so it is possible to adapt the unification process to the characteristics of the texts.

Another important aspect is that the processes of word-joining, word-splitting and common-letters are executed both together and in different phases. In a first phase the new words are required to be the same as some of the already existing words. In later phases this restriction is eliminated, and the new words need not to be exactly the same as the already existing words. There is a limit, however, the ratio of the number of identical to different letters must be equal to or greater than 1. This loop is executed whenever there are positions left in the unified version without a word assigned to them, and the previous processes give positive results.
Finally we must point out that the user has the control of the execution of the previous process the same as he has for the unification of positions, this is to say, deciding if it is executed or not and the rank of positions in which they work. An interactive subset has been added in the MS-DOS version, UNITE91, allowing the user to interact with the computer during every moment of the collation process. Like every other routine inside UNITE, it is the human editor who decides whether this enhancement will be applied, how, when, and where.

The automatic unification process generates two output files. The first stores the unified version, while the second enters data relevant to the three previously mentioned processes and sums up its execution. The collated texts are also divided into stanzas, each separated by a blank line. In addition to the words selected for the unified version, there are also the variants not selected, with a number assigned for the identification of the original version where they appeared. Optionally, the original stanzas of all versions may also appear preceding the unified stanza.

The file that sums up the process of unification is created optionally and controlled by a user-modifiable parameter. Its generation is very useful whenever an explanation is needed of the execution of the aforementioned of the processes. For each word generated by those processes, a file line indicates the stanza and line in which it appears, the process that generated it, the unified word resulting from the process and the original words (accompanied by a label identifying the corresponding version) that generated that unified word. These words appear with their original spelling and not with the unification processes work.

TACT

TACT is a program for text information retrieval produced by the Centre for Computing in the Humanities of the University or Toronto in Canada. The basic way of working of TACT is quite easy: it allows to ask about where a group of words or phrases are. To do that, the program does not read a textual file, it works with a Data Base which is made by one of its tools, MAKBAS.

The user can see the text, get a list of words, choose the words to show, in order to build up concordances or referential index of smaller size, study the relative distribution in the text with statistic aims. The texts can be written in any system of characters, TACT allows the redefinition of the system through an auxiliary file (XLATTABL.DAT) or using combinations of keys. The program works on selections, with the advantage that the user can establish a exclusive control character and build up a data base in order to obtain specific information.

The user can personalize his or her wishes through the creation of personalized rules or through the personalization of the textual base which can be kept for a later enquiry. It allows wide margins in the definition of contexts or in the selection, even in the concrete application. The data base can be interrogated not only about what context a word appears in, but also with more complex consultations, such as the number of words ending by -ed, or the places where a character informs about a specific word or its formal variation through the use of substitutions, or where in the same paragraph a series of words are used.

The user can build a list of synonyms through some rules: alterations of the verbal paradigm, for instance, and search through them. These rules can be imported, exported or kept. The searches allow conditions, positive or negative, combinations of requirements and substitutions. The obtained results can be printed or we can kept in
disk, in ASCII format, replacing the last file with the same name or being added to it, that depends on the choice selected. Another way of using the work of other sessions is by creating a file SCRIPT, containing a register of already made operations which can be repeated in any other session and that have been individualized under a proper name.

TACT works with other programs like COLLGEN which reviews the words in a text with the purpose of finding all the places where a combination of two or more words appears more than once, and shows a list of combinations and how often they appear. It also has the advantage that it can be processed by sections. The major advantage of TACT is the possibility of creating its own textual data base and defining all the parameters that will be needed.

The textual data base contains not only the text, but also a completed index of the whole position of all the words in the text, information about the formal structure: where a chapter of a novel or a scientific book starts or finishes or where a character starts or stop talking in a comedy. MAKBAS is the program which provides all this kind of information, together with a previous labelling, the characters and used diacritic signs of those parts which are not included in the text itself, such as notes, references or complements, all of it from files in ASCII format. Sometimes we may try to create a textual data base too large for MAKBAS. TACT solves this difficulty by using MERGEBAS, a program which allows to combine several textual bases in a larger base, solving the inconvenient of an operative system of small memory resources, as MS-DOS.

**Image processing for manuscripts and incunabula**

The digitization and image techniques began to be developed and commercialized by the North America Spacial Agency, NASA, in the seventies, allowing the conversion of an image in a digitized standard. They started as tools used in the natural and biological sciences; nowadays, they have moved to the humanistic field. Because of the increasing of memory capacity in PCs and the rise of the possibilities of storing the large files resulting of the digitization, in the very last years, this technique has approached to the user, so that compact-disk ROM has allowed a definitive reduction in the price of the costs, and the possibility that everybody can benefit from it in the personal research.

The digitization of manuscripts and incunabula offers two essential features: it preserves bibliographical heritage and provides facsimiles far superior to those available through conventional photographic techniques. The collaboration of the Biblioteca Nacional in this process is essential: it guarantees that the work will be carried out under conditions of maximum security and without damage to any of the volumes being digitized.

After many studies it has been found out that a previous step to the digitization of the manuscripts is needed: the obtaining of colour slides of the original, which also allow to improve the conditions of reading, by the use of the most suitable auxiliaries (ultraviolet and infrared light, for example). In those cases where digitization from slides was not convenient, direct scanning of the original volumes has been performed.

By carrying out the previous step several advantages are obtained:

- There is a minimum physical contact with the document.
- The document is out of its usual place a very short time.
- The document is exposed to the light only during more or less 15 seconds per page.
• There are not strong changes of temperatures or humidity.
• The books do not have to be totally opened, so the bindings are not forced.
• The processes of digitization are carried out using the photographic images which allow electronic digital images to be obtained without the presence of the originals.
• It is possible to explore or study very degraded pages without reprocessing the originals ones.
• The Biblioteca Nacional can use the colour photographic archive for scientific purposes.
• The legibility of some documents can be increased by combinations of light, filters and emulsions sensitized in a special way. It is possible to obtain better results than with other techniques.

Once the technique to be used was decided, we selected the photographic material by choosing a reversible colour film of 24 x 36 mm of very fine grain with a maximum resolution. The digitization of the photographic images of the pages of the book is the bridge between the photographic and the electronic processes used in ADMYTE. In order to carry out the phase of digitization we have used slide scanners of high quality and maximum resolution (4,096 points per inch, 3850 x 5800 points in each slide) which are able to differentiate 16 millions of colours. To carry out the storage of those images into WORM Optical Disks, it has been necessary to develop new algorithms for the compression of colour images.

The procedures of image processing allow to make the documents more legible, taking out stains of humidity, yellowish colours, natural attacks etc. The procedures include statistical methods for reduction of colours to their medium value, control of tones in different places, substitution of colours and dots, and relief of contours and contrasts, which let obtain spectacular results when the background of the page is cleared up and the inks darkened.

The processing of images concludes when arriving to black and white pictures and their reduction to their equivalent to 150 dots per inch. This process allows us to obtain a printed copy of high quality on which we base the verification with the originals. Once the quality is obtained, a mechanism of compression is carried out having as a result an average of 39 KB per image. Those parts which contain illuminations or pictures which will be convenient to reproduce in the final database receive a special treatment. For these, special programs have been created which allow to carry out the processes of correction in an interactive way, this is, displaying in the screen the results obtained after the modifications.

Electronic transcription of digitized incunabula

Scanners and text recognition technologies have improved so much in the last years that nowadays they are commercially used because of its high degree of confidence. Although the accuracy of reading is not perfect, we can say that the results are so good that we can think of the immediate disappearance of the manual transcription of printed books in the modern era.

The first printed texts, the incunabula, are not usually clear enough for electronic reading and the later automatic digitization. The accomplished studies make us think of the possibility of carrying out the electronic transcription though scanners. It would not be necessary to submit the incunabula to a process of digital reproduction which will damage their bindings. ADMYTE develops a technique which is not based on printed texts, but on their reproductions in colour slides, direct scanning will be made only when it be the only possibility.
Producing digitized images from celluloid film has been already been successfully applied in art history and in documentation. The development of this technique to the concrete application on librarian science will be a remarkable advance in the service of libraries.

The application of OCR techniques (optical character recognition) is not essential to ADMYTE. The transcription of texts, however, will benefit from the cooperation of all the scholars who would like to give their transcriptions in electronic from, and, from the possibility of reading optical edited texts and typewritten texts, we will reduce the need of typing the texts for data introduction in electronic format and we will improve noticeably the quality of the ASCII texts included in the disks, because of the corrections of the editions or mechanic transcriptions read electronically. There is no doubt that the quality of the text transcription is an essential requirement in ADMYTE.

CLARITY-CD and WINDOWS

ADMYTE allows the use of a personal computer (with a high resolution VGA screen), a CD-ROM driver and a laser printer for immediate and safe reproductions. An ad hoc version of CLARITY-CD is developed to combine all these elements and take advantage of all the information contained in them. CLARITY-CD is MICRONET’s retrieval program for texts and images that can support a great amount of peripherals, scanners, CD-ROM drivers or printers, as well as several screens. Two kind of indexes, GLOSARIO and INDICES.DEC, can be used simultaneously and with multiterm descriptors to extract the information required in the data base. Ten queries can be done simultaneously, although macro-query can provide batch processing facilities with a reorganization of the extracted information. The queries use a collection of Spanish and English logical operators that allow a very extensive combination.

When searching, CLARITY-CD not only works with words but can also accept request of complete sentences of more than forty alphabetical or numerical characters: it makes easy the search among dates and is able to distinguish ranks among numbers. Three search environments have been developed aiming to the application of CLARITY to texts included in ADMYTE: hypertext and hypertext are joined to the traditional context. It allows the establishment of a range more or less wide, delimiting whether the area of searching is the whole text or only a parcel of it, defined by the user.

CLARITY combines the possibilities of a data base with records of variable length with the advantages of having its own editor with text processor and a system to import information from other data bases, without losing any relevant element of that imported information. It is presented like in a Windows 3.0 environment that allows to compare text and image on the same window or in cascaded format, to watch the structure of the base being worked with, to select in a simplified way, by glossary or enlarged, which is the most complex, to keep the queries in a disk by means of macros of queries that are stored and can be used again to create new data bases with a limit of sixty-four fields and a length of thirty-two characters in each definition, and to associate a key word to the bases to protect the information. Other windows allow the use of the text processing system. That is flexible enough for an adequate presentation of the reports obtained from the information on the base, without turning to other system of texts processing.

The list-window allows not only to list by printer, but to save a file in a disk with format, to save it in a disk as an ASCII file or to save images in files with the standard format of storage and interchange of images Tiff or compressed Tiff. In the second case, it needs less space on disk.
Conclusion

ADMYTE is a project that combines a lot of elements of different types and origins: public and private centres, teaching and research institutions, and libraries, investigators with humanistic and technical background, all of them within a Spanish framework, but coming from different countries, languages and continents, in the service of one aim: a collection of Old Spanish texts in CD-ROM format. MICRONET will propagate this collection through the usual commercial channels, so that the funds that ADMYTE contains will be available all over the world. ADMYTE is the largest effort of spreading the graphic cultural heritage (texts and images) carried out by a country so far.

The consequence of this effort is double: in the policy of culture, as it has been said, it is the major aim so far, to spread out the cultural Spanish heritage in a global way; in the field of technology is an answer to the need of making Spanish a language of the 21st century, showing how it has been managed in applying the most modern resources of text and image processing and retrieval to the service of the cultural heritage, in a project carried out in Spain by Spanish firms and institutions, with international and inter-university collaboration.

Bibliography


