Computer Generated Analytical Indices for the Zuozhuan

John Page [1], **Isabel Garcí a Hidalgo, & Rosa Elena Moncayo,**
El Colegio de Mexico, Mexico

This project is designed to create English/Spanish-Chinese computer indices to facilitate analysis and translation of the late 4th century BCE *Zuozhuan*. The project is financed by Mexico's National Council on Science and Technology (CONACYT).

By data entry of the Chinese text, without intercalated commentary, and of the Chinese-English lexicon entitled *Index to the Zuozhuan* by Fraser and Lockhart (FLI)[2]and linkage of the two in a database, it is proposed to complete and up-date the FLI as a tool for translation and create English-Chinese and Spanish-Chinese subject indices for the *Zuozhuan.*

Description of the text: The *Zuozhuan* is a fourth century BCE historical text which narrates events during the so called Spring and Autumn period (-722-468 BCE) of the Zhou dynasty inconclusively attributed to Zuo Qiuming, a contemporary of Confucius. The canonical category of the text is based the fact that it has come down to us as one of three commentaries on the *Spring and Autumn Annals* of the feudatory state of Lu, where Confucius was born and to whom its compilation is attributed. It has thus been studied, edited and commented upon since the Han dynasty.

The *Zuozhuan* is composed of 179,570 graphs, in a consistent literary Chinese of its period, edited to conform to the Annals of the State of Lu. It contains a wealth of political, military, social, geographic and astronomical information as well as literature, philosophy, folklore, religion and popular belief shared by the inhabitants of Lu, sixteen other Zhou dynasty feudatories and a great number of other smaller ones absorbed during the period by larger more powerful neighbors.

[1] John Page, Professor of Chinese Literature, project author and director; Isabel García Hidalgo, database and system completion and implementation; M.Gerszo, original database design, data entry and printing supervision and implementation; R.E.Moncayo, F.Haro, Chinese data entry.

[2] Fraser, Everard D.H. and James H.S. Lockhart, *Index to the Tso-chuan* Oxford University Press, 1930, reimpresiòn, Ch'eng Wen Publishing Co. Taibei, 1966.

No extant text contemporary with or prior to the *Zuozhuan* contains as much and as valuable information about this period in the history of China. However, while due to certain characteristics it is a difficult text to study, these same characteristics make it appropriate for computer analysis. Thus, for example, not all the feudatories mentioned in the text, are consistently covered every year, nor are all the events in all the feudatories presented every year, thus considerably complicating continuity and coherent reading of the text. Important actors are not referred to consistently by a single title or name but by other appelatives or names, social styles, political and posthumous honorific titles which, as was traditional in early China, could change during the individuals lifetime and a narrative of that life.

Over a period of two thousand years, since the beginning of the Han dynasty, a vast and valuable corpus of commentary and explication has accumulated with respect to the content of the *Zuozhuan*. This is customarily intercalated in the columns of the printed text, further complicating its reading in Chinese and in translations into Western languages which follow the same procedure. To these difficulties must be added the density of the language. Its universe of 179,570 graphs condenses to a net lexicon of 3,789, setting aside 1,214 graphs that register one meaning each, the reminder share approximately 15,491 meanings, including titles and the names of persons and places.

The FLI: The Chinese characters in the FLI are organized solely "by radical" and additional strokes, that is by a series of 214 elements visible in those characters which were chosen in the early 18th. century as classifiers and, within each classification, by the number of strokes additional to the radical. Since the phonetic value of each graph was registered in the FLI in a transliteration no longer in use in English, that system has been replaced by Wade-Giles and *Pinyin*; our database will allow search by both methods. The FLI compilers did not include all the meanings of all the characters and all the positions they did register. In spite of the fact that only 242 graphs are missing in the FLI, many meanings of graphs that were registered are lacking, and these fall into two categories. The majority consists of grammatical indicators, which in the light of their category vary little; the second consists of very frequent lexical characters, in which case, the compilers included many meanings but indicated the remaining repetitions, and among them further variants, by such Latin terms as *et saepe, etc., et alia, passim,* in the sense of "there are more of the same."

Early literary Chinese does not include many true digraphs in spite of the fact that many names of person and places are composed of two, three and four graphs. The FLI rarely presents those names complete, but rather requires the reader to follow cross references to find the other characters included in the name or the title of the individual or place. This solution, in all probability, was aimed at reducing the bulk of the FLI and, thereby, the cost of publication, for which there would have been a very limited market at the time.

None of the positions of characters in the FLI includes a line number. Only page and column are specified in the printed text which is set at 18 columns and 42 lines per page, though not uniformly.

The project: Fortunately, at the outset of the project, a Windows compatible computer program for Chinese was already available facilitating data entry of the Chinese characters in the text and the lexicon by means of alphanumeric codes and a tool with which to create characters not included in the software dictionary.

The printing of the text chosen is a punctuated version without intercalating commentaries of the 1815 edition by Ruan Yuan (1764-1849), a distinguished Ching literatus, and the edition used by James Legge, its first translator into English.[3] The FLI is based on this edition which is included in the established texts of the thirteen classics (*Shisanjing zhushu*).

A text in literary Chinese formatted for traditional printing appears as a grid of characters in parallel columns and lines to be read from right to left and top to bottom. Each character occupies a cell in the grid. The meaning of a given character may thus be registered and obtained according to the intersection of any column and line on any page. The FLI as published indicated only page and column, our up-dated version will not only permit a search by column and line on any page but searches by character will locate the full coordinate. Thus the reader or translator will obtain the meaning of a character at any intersection on any page specified. The database will also permit combinations of two, three and four graphs and will differentiate between different meanings of the same character repeated in the same column, or show all the meanings of a graph in the case of a general search.

Data entry, 1) The problem that immediately appeared was the need to digitize Chinese characters together with meanings in English, not only of the entry characters, but of the additional characters which, combined with the entry character constitute names of persons and places. The task was carried out by two different teams: one with no knowledge of Chinese, and the other composed of two M.A. level sinologists. The first team recorded meanings in English, data for page and column, adding a slash in the position of each

[3] Legge, James. *The Ch'un Ts'ew with the Tso Chuen*, Oxford University Press, 1895: reprint, Hong Kong, Hong Kong University Press, 1982.

Chinese character found. The Chinese characters were digitized by the BIG 5 alphanumeric code number used for characters written in the traditional manner without simplification of strokes, and included in the Chinese software *Twinbridge 3.3*. Since the *Zuozhuan* includes archaic and relatively unusual characters not included in the *Twinbridge dictionary*, these were drawn using the *Twinbrige editor*, thereby increasing the net number of characters to 3,909. At the same time, it was necessary to place 33,698 characters among the meanings in English. Errors immediately appeared in the number and placement of the slashes indicating where the Chinese characters were to be added, as well as other errors characteristics of data entry.

Data entry, 2) The *Zuozhuan*: Since one of the principal goals of the project is to take full advantage of the FLI, linking the database with the Chinese text, the first step was to preserve the page and column information as it appears in the published Legge printing so as to subsequently add line numbers. This was achieved by digitizing the text directly into a grid that precisely reproduces the number of columns and lines on the page, creating a *Word 6* file for each page. Though the number of columns per page varies little (18), the number of lines varies greatly (from a maximum of 42 in most cases, to as few as one). The complete text has been successfully printed.

The computer system: When proposing the use of the computer for analysis of the *Zuozhuan*, the following objectives were established: 1) The definition of a method that would permit association between every occurrence of a character in the *Zuozhuan* with its meaning in English (and subsequently in Spanish) or any other characteristic of interest for analysis or translation of the text; 2) The availability of selective recovery procedures for those meanings and characteristics; 3) The creation of user-friendly resources to handle Chinese characters by means of controls like those used in *Windows*; 4) The creation of a database in which Chinese characters were represented by alphanumeric codes would have fulfilled the first and second objectives but not the third. To fulfill all the objectives established, it was necessary to create a database that in conjunction with other resources would facilitate simultaneous management of Chinese characters and characters in the English and Spanish Roman alphabet.

The first computer solution

A system was designed for *Windows PC* consisting of the following three elements:

1) The *Zuozhuan* text organized into 351 files, each of which correspond to a page of the Legge printing of the Ruan Yuan edition. Each page/file is stored in the *MS Word 6.0* format. The information in these files is displayed by means of *Twinbridge Chinese System* 3.3, which converts alphanumeric codes into Chinese characters.

2) A relational database developed in *Personal Oracle 7*, a database administrator that uses the *SQL (Structured Query Language)* standard. This database mainly stores the information in the FLI and the Chinese characters handled by alphanumeric codes.

3) *Word* macros, written in *WordBasic*, by means of which the Chinese characters are appropriately managed and the selective recovery and/or up-dating of information in the database is carried out. The user controls the execution of these macros by means of buttons in a *Word* tool-bar specially designed for the purpose. The macros utilize extensions of *ODBC (Open DataBase Connectivity)* library routines for *Microsoft Word* to access the database. Further macros required by the system would be designed and programmed as the project proceeded.

In this solution, the three above mentioned components: relational database, text in *Word* format, and macros for specific purposes that interconnect the text with the information in the database, in fact constitute an information system the interface of which to the user is a customization of *Word*.

The foregoing first solution has evolved to the point described below:

1) The text of the *Zuozhuan* has been completely digitized, each page of text being stored as a table so that any character's location is defined by the coordinates file-page, column and line. The computer text has been checked and corrected, and printed as it appears in the Legge original by means of macros that eliminate the table format used for storage.

2) A database was designed to represent the FLI and contains five related tables (see figure 1) as described below:

CCHAR: contains a row for each entry character in the FLI. Data for each entry character include: name (phonetic value in Wade-Giles transliteration), FLI entry number, BIG5 code, *Pinyin* transliteration, FLI number and BIG5 code for the radical, number of additional strokes, and a unique identification number, CCID, which relates this table to the table ASOCIA.

ASOCIA: is an intermediate table whereby any entry character in CCHAR relates to its corresponding meanings stored in the table SIGNIF. It contains a row for every meaning of every entry character in the FLI. These rows include the CCID identification number of the entry to which the meaning belongs, the unique meaning identifier, SIGID used in SIGNIF, and its own unique identifier ASOCID. This organization makes it possible to recover all (one or several), meanings or any meaning of an entry.

SIGNIF: Contains a row for every meaning of every entry character in the FLI. The rows are uniquely identified by their SIGID value. The meanings in the FLI combine words in English with Chinese characters. The database cannot easily handle Chinese characters interspersed with English words, hence, a slash is used to mark the location of every Chinese character within an English phrase. Each slash is recognized and replaced by the corresponding BIG5 code by means of SQL instructions that query the database, controlled by the *WordBasic* macros. These macros display the results in a *Word* document in which the resulting BIG5 codes are interpreted as Chinese characters by means of the *Twinbridge* system. Since it is possible to find several Chinese characters in a single English phrase, the macros utilize location information stored in the CCCODIGOS table.

CCCODIGOS: contains a row for every Chinese character located among English meanings including the hexagrams and trigrams of the *Yijing.* The information stored in each row includes the character represented by its FLI number, the page of the FLI in which the character is located, the FLI number of the entry to which the character belongs, and a unique identifier with which the order of appearance of the Chinese characters throughout the meanings of an entry is controlled.

OCUR: This table relates the meanings contained in the FLI with the characters in the *Zuozhuan* text. The table contains a row for every Chinese character in the text indicated in the FLI. Each row contains the location (page and column) of a character in the text and the unique ASOCID identifier used in the ASOCIA table, by means of which that character may be recovered in its context.

The structure and organization of the database reflect solutions for the simultaneous management of Chinese and English designed to facilitate the stage of digitization of the information and to be used at the stage of recovery.

The information in the FLI was digitized into the database until it contained 3,909 rows in the CCHAR table, 16,705 rows in the SIGNIF table, 33,698 rows in CCCODIGOS and 93,056 in OCUR.

A macro was developed for recovery of meanings appearing in the FLI for every Chinese character in the *Zuozhuan* text. Furthermore, another macro was designed, which, utilizing the location data of the characters in the text, would make it possible to incorporate the missing coordinate, the line, into the OCUR table in the database. So that, the locations and meanings of the 86,514 characters in the *Zuozhuan* that were not indicated in the FLI could be processed and incorporated into the database.

The current computer solution

The information in the FLI was digitized but not checked and corrected where required. It was necessary to program macros to facilitate the checking and correction of that information. In fact, this was the beginning of the stage of design and development of macros that would make use of the information stored. It was at this point that the decision was made to make access to the computarized *Zuozhuan* and FLI available to other researchers on the *Internet*. That decision led to the computer solution now in use and development.

The present system consists of three components: the database, the interface to the user, and a connecting module between the two:

1) The database representing the FLI now resides centralized in a server of the Colegio de Mexico's local network. This database was developed by means of an administrator capable of handling client-server architecture, a *Microsoft SQL* server for *Windows NT***.**

Under this scheme concurrent queries to the data base may be made from any node of the *Internet Network*. The information from the first database, the database in *Oracle*, was transferred to a database defined in *SQL Server*. The essential structure of the *Oracle* database was preserved since both the *SQL Server* and *Oracle* conformed to the *SQL* standard.

2) The interface to the user is constructed in the shape of documents and forms written in *HyperText Markup Language (HTML).* These may be displayed on the computer screen by means of "browsers," such as *Netscape* or *Internet Explorer* on the *World Wide Web (WWW)* of the International Internet Network. Users execute queries or up-dates to the database by means of buttons with specially programmed functions included in the forms exemplified in figure 2.

3) The system utilizes the component of the *Microsoft Internet Information Server* called *Microsoft Internet Database Connector* (*IDC*) to execute queries to the database and receives the results in the form of an *HTML* document. *IDC* uses *ASCII* files with the ".idc" extension in which the *SQL* instruction (to consult the database) that it is wished to execute has been programmed, as well as the name of the file containing the form of presentation in which the results of the query will be presented to the user. These last files carry the extension ".htx" and are similar to documents in *HTML* except for two characteristics: They contain key words that control the form in which the results of a query are to be presented to the user, and contain markers which specify where to include the results of the query in the *HTML* document.

The documents and forms written in *HTML*, the ".idc" files, the *SQL* coded instructions included in them, and the information relating to the ways of presenting the results coded in the ". htx" files constitute the computer tool for programming the functions that, in the first solution, corresponded to the group of *macros*.

In the present system, the interface to the user is composed of *WWW* pages: the documents or forms written in *HTML*. Representation of Chinese characters appearing in the *WWW* pages is still executed by the *Twinbridge* system. However, the network environment allows the interface to the user to follow the "client's" configuration. For the time being, we consult the database on the *WWW* pages in either the *Windows 3.11* environment with *Twinbridge 3.3*, or in the *Windows 95* environment with *Twinbridge 4.0*.

To achieve complete functionality of this system, we will change the *Zuozhuan* text from its present format as *Word* files to a table in the database appropriately related to the tables representing the *FLI*. Procedures will be programmed to allow the representation of the text by segments that may be the original pages or the annals by which the text itself is organized.

The *Zuozhuan* system in its present form is still under development, but we are in a position to show some of its results. Figure 3 shows a page of the *WWW* presented by the system when an entry in the computarized *FLI* is consulted. It displays the Chinese characters integrated into their corresponding English meanings.