RELIGIOUS TEXTS
Alfonso de Cartagena. *Oracional de Fernán Pérez* (Murcia, 1487-3-26)
Alfonso de Cartagena. *Contemplación sobre el salmo judgame Dios* (Murcia, 1487-3-26)
S. Juan Crisóstomo. *Tratado que demuestra que ninguna persona se daña sino a si misma* (Murcia, 1487-3-26)

Disks 3 and 4 are partial "reprints" from Disk 1, with the transcriptions and digitized facsimiles of the *editiones princepes* of the *Siete partidas* (Sevilla, 25 October 1491) and of the *Libro de proprietatibus rerum* of Bartholomaeus Anglicus (Toulouse, 18 September 1494), respectively.

OMNIPAGE AND *WORDCRUNCHER*: TOOLS FOR CREATING AND SEARCHING DIGITALIZED TEXT CORPORA

Mark Davies
Illinois State University
e-mail: mdavies@rs6[illegible]p.iistu.edu

With the recent release of the ADMYTE CD-ROMs of Old and Middle Spanish texts, researchers in Ibero-Romance linguistics and literature now have available a large and useful corpora of texts for analysis. However, in addition to analyzing the texts on these CD-ROMs, there also exists an alternate means of obtaining large electronic corpora of texts: researchers can generate their own electronic texts for analysis by using a scanner to input the texts into the computer and then indexing them to permit quick, yet complicated searches of the data. In the past five years I have scanned and indexed over 5,000,000 words from Old, Middle, and Modern Spanish and Portuguese texts, taken from 95 books and representing 154 individual authors. The texts have been scanned into a PC using *OmniPage Professional* software (Windows versions 2.11 and 5.0) for optical character recognition ("OCR"), and indexed and analyzed with *WordCruncher* text analysis software (DOS version 4.5). This review considers some of the strengths and weaknesses of both products, and gives as well some general suggestions for creating one's own electronic corpora.

Selecting the text

To produce electronic corpora for analysis, a researcher must first acquire suitable texts for scanning. The purpose of the OCR package is to convert a graphical image of the scanned page into individual characters, which can then be loaded into a word processor or text analysis program. Therefore, one important consideration is that the printed page have a standard typeface. For example, it should not be overly ornate, as is the case with many books printed in the earlier part of this century. It should also have adequate spacing between letters and not contain typeface with potentially confusing letters. For example, in several books originally chosen for my corpus, the overly small space between the stem and the diacritic of the "í" ("i" with acute accent) caused the character to be recognized repeatedly as an "f". The darkness of the text also should not vary considerably from one page to the next, because the scanner will then have to adjust for this difference from page to page.

82

The binding should not be so tight that it prevents all of the text on both facing pages from lying flat on the scanner bed. Also, the smaller the characters, the more important it is that they not be too dark, that they use a standard typeface, and that they not run together. Finally, numerous footnotes in the text are disadvantageous, because it will be difficult to retain the footnote structure in the electronic text, and the footnote numbers in the text itself may be scanned in as part of the word they follow.

Scanning the text (with *OmniPage Professional*)

Assuming that one has a "clean" text to work with, the next stage is to scan it. *OmniPage* has certain benefits and disadvantages, but has several features that make it well suited for scanning texts from older stages of the Romance languages. There are two main advantages of *OmniPage* over several other OCR packages. First, unlike some products (such as the popular *TypeReader* OCR software), *OmniPage* can be told which language is being scanned, and mix together several languages. This is rather useful in Old Spanish texts, because one can tell *OmniPage* to accept both Spanish and French characters, so that the "ç" is recognized as one character, and not "c" with a comma, which would probably happen if only Spanish (i.e. Modern Spanish) were chosen. The second advantage, one that makes *OmniPage* ideal for Old Romance texts with non-standard characters, is its well-developed "training" mode. If the text contains non-standard characters such as the "et" sign of Old Spanish or the nasalized "i" and "u" of Old Portuguese (i.e., "i" and "u" with a tilde), *OmniPage* can be trained to recognize these characters as such, rather than simply as an unrecognizable character (such as "et"), as the character alone ("i" and "u" without a tilde), or as the character plus separate diacritic ("i~" and " u~"). For example, even though *OmniPage* does not include Old English as one of its twelve recognized languages, I was able to train it to recognize the eth (ð), thorn (þ), and ash (æ) as such, in scanning the Gospel of Luke in Old English. This training mode is useful not only for standard characters from the alphabet of a certain language, but also for characters that are consistently difficult to recognize throughout a text, as in the case of the confusion between the "í" and the "f" noted above.

Regarding the amount of time required to scan texts with *OmniPage* and the difficulty involved in doing so, I offer the following comments. In a book with good quality typeface and the other characteristics noted above, I have been able consistently to average about 150-250 pages (75-125 scans of facing pages) per hour. Factors that affect the scan rate are, of course, the quality of the printed text, the configuration of the computer, and the physical layout of the pages. Concerning the computer: a 486DX2-50 with 8 MB of RAM should be sufficient, and users of Pentium machines will obviously notice increased speed. Concerning the page layout: most researchers will want to eliminate from the electronic corpora all footnotes and page numbers. They can do this at the scanning stage, by choosing to "recognize" only the body text. It does take more time to identify the body text in each scan, but *OmniPage* can be set to recognize the same portion of the page from scan to scan, if the page numbers and footnotes are always in the same place. Alternatively, one can have *OmniPage* recognize all the text on the page, and then manually edit out all unwanted material with a word processor.

The steps required to set up *OmniPage* for scanning and then to make multiple scans are not complicated. Before scanning the first page, one has to identify the language of the text, the type of scanner, and the size of the page (legal, letter, A4). One also decides whether to scan in multiple pages and then save all of them together (the fastest option), or whether to save page by page (very slow, but the safest option). The user also chooses to have *OmniPage* automatically adjust for the darkness of the printed text (the slowest option), or can choose to set the intensity at the beginning, and then manually change the setting for especially light or dark pages. *OmniPage* provides the advantage of being able to change any of these settings between individual scans, which some other packages do not allow. Unless the page darkness changes markedly from page to page and one is forced to adjust manually for this (i.e., automatic brightness adjustment is not chosen), and unless one decides to identify manually the regions for OCR on each page, then *OmniPage* can often recognize the contents of the page in just a bit more time than it takes to flip to the next page. Once the pages have been scanned, *OmniPage* saves the contents to disk in one of several wordprocessing formats, including ASCII, RTF, WordPerfect, Word, etc.

The next step is to edit the texts. I have found WordPerfect 5.1 for DOS to be ideal, because of its ability to handle large files (1-2 MB), its speed vis à vis Windows word processors, and its simple macro programming, which greatly simplify repetitive search and replaces. For example, I have created a macro that looks for characters that *OmniPage* was unable to recognize (usually replaced by a "~" sign), and then corrects them in one of several ways, depending on a one or two keystroke sequence.

Analyzing the text (with *WordCruncher*)

To perform complicated proximity and Boolean searches on text data, it is necessary to search the corpora with a text retrieval and analysis program. There are several such products available, but I have found *WordCruncher* to be the most powerful, at least for my particular needs.

In order to search the texts with *WordCruncher* (using its "WCView" module), the texts must first be indexed by *WordCruncher* (using its "WCIndex" module). To prepare the texts for indexing, one needs to insert codes to identify page and paragraph divisions, although *WordCruncher* can often do this automatically. On a 486DX2-50 PC, *WordCruncher* can index a file of 500 KB in just over a minute. The indexing program creates a frequency count of each individual word in the text, but more importantly, creates an index of the location of each individual occurrence of every word in the text.

Once the every-word index exists, the WCView module can search the texts. Its fast yet advanced search capabilities are one reason that *WordCruncher* has gained some measure of popularity in the humanities computing community. Consider an example from my 1,650,000-word corpus of Old and Middle Portuguese texts: *WordCruncher* searched for every instance of the object pronouns "me", "te", " nos", "vos", "lhe", and "lhes" (25,000 cases) immediately followed by any forms of the verbs "poder", "dever", or "querer" (16,000 cases). All of the forms of the verbs had previously been identified, and a file had been generated that contained the location of all of these 16,000 cases. The ability to save search results in files and later to combine them is one of the most useful features of *WordCruncher*. Once the pronouns and verbs were selected, *WordCruncher* was able to find all 1000+ cases of a pronoun immediately followed by one of these three verbs in just two-tenths of one second. *WordCruncher* then searched for all cases of these 1000+ constructions in which they were immediately followed by any infinitive (57,000 cases). This search, which resulted in 689 "hits" (on phrases such as "lhe quero falar", "me devem pagar", etc.) took another four-tenths of one second. Besides simple proximity searches of the type just described, one can also perform more complicated Boolean searches that include multiple combinations of "and", "or", and "not".

Once all the relevant examples are found, *WordCruncher* can also provide distributional information, including the number of examples in any "range" of dates, topics, or texts previously identified by the user (e.g. "1200s", "religioso", *El Corbacho*, etc.), and compare this against the expected number of examples in those ranges. The examples can also be sent to a printer or to a file, either as a list of references only, or with the examples preceded and followed by any given number of lines of text. In addition to these searching capabilities, of *WordCruncher* possess many other features, such as the ability to create different types of concordances.

Conclusion

In summary, there already exist various powerful computer tools that permit researchers to create their own electronic corpora of texts. There are a number of OCR and text retrieval and analysis programs in addition to the two discussed here–*OmniPage* and *WordCruncher*. Depending on researchers' interests and resources, one of the competing products may be more suited to their particular needs. Regardless of the specific package used, however, a researcher can, in a matter of one or two days, scan a book-length text and then perform complicated searches on it. As humanities researchers become more skilled in creating and disseminating electronic corpora, the quantity of computer-based resources available to us as an academic community will expand significantly.

Product Information:

OmniPage Professional Version 5.0 is available directly from Caere Corporation (1-800-462-2373) at a list price of $695.00, but also from software distributors at discounted prices around $400.00

WordCruncher is available only from the publisher, Johnston and Associates (1-812-339-9996), which offers an academic discount price of $299 for DOS version 4.6 and $399 for a new Windows version. The Windows version is still in development, and at present lacks some of the more powerful search features of the DOS version.