DOCUMENT RESUME

ED 046 296 FL 002 086

AUTHOR Kelly, John P.
TITLE Portuguese and the Computer: "uma bossa nova".
PUB DATE Dec 69
NOTE 15p.; Paper presented at the meeting of the Linguistic Society of America, Chicago, Illinois, December 1969

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Character Recognition, *Computer Oriented Programs, *Computers, Computer Science, Educational Technology, Information Processing, *Language Classification, Language Patterns, Language Research, Literary Analysis, Literature Reviews, *Luso Brazilian Culture, Optical Scanners, *Portuguese, Romance Languages, Word Lists

ABSTRACT

This paper describes the process of preparing a computer study of Brazilian Portuguese literary texts to be used both in teaching and in the preparation of a reference text. Procedural difficulties encountered in the project point out the potential and limitations of computerized research in literary studies. Seven possible areas for computer applications are outlined with reference to specialized texts for further study. These include the preparation of word lists, bibliographical lists, and concordances; linguistic and content analyses; attribution studies; and critical editions. (PL)

"PORTUGUESE AND THE COMPUTER:

UMA BOSSA NOVA"

John R. Kelly

April of 1968 marked the beginning of my first timorous steps in Computerlândia. My motivations for undertaking the initial project were 1) To have an updated frequency and range list of Brazilian Portuguese for use in my own classes and for possible publication as a reference text; 2) To learn by my limited study something more of the possibilities of the Computer in the area of language and literary studies. I was influenced to begin my word list because of successful research in progress or completed by colleagues in International Relations, Classics, and English at other campuses. There was, of course, a certain amount of seduction involved by the winking and blinking mystery machine at the Computer Center which stood on my daily route to the Library.

The mechanics of my project went in supposedly logical sequence. I first requested and was alloted a certain amount of computer money which could be used for card keypunching, programming charges, and real machine time. The Computer Center director was enthusiastic in his encouragement of my proposed word study. He wanted the Humanists on

our campus to cease dragging their feet and to cast off their prejudice against his unfeeling electronic adding-machine. He warned me that I must not become discouraged. This was to be an important piece of advice, and one to be remembered by any who would contemplate entering into computer assisted projects. Second, since I am not a programmer and it is only now that I have some rudiments of programming, I had to seek the help of one of these men. I had to explain my proposed project to him so he in turn could devise a way of getting my Portuguese words into the machine and some ordered results out of it. In my innocence, I assumed this was going to be the easiest part of my work, but communicating with that man became tantamount to a Herculean Labor. As an example, the programmer mentioned language in his conversation, I naturally believed, being the liberal arts man I am, Portuguese, Spanish, French, English maybe... He, on the other hand was referring to assembly language, machine language, SNOBOL, COBOL, FORTRAN, and the like. He immediately recommended SNOBOL for manipulation of "data strings", adding sadly that no one on our campus knew that language. After a total of ten minutes of conversation, he did deign to consider my word project. He suggested, as the next step, that I get my texts in machine

readable form. This meant, in this instance, punched cards. While he was left to ponder the intricacies of the Portuguese Alphabet, I went on to the key puncher.

Conversing with the key puncher was nearly as ethereal as my chat with the programmer. I explained to the girl that I would be bringing her samples of written Brazilian Portuguese containing approximately 2000 words. The samples would be from novels, verse poetry, magazines, newspapers and the like. She stared silent, sullen and uncomprehending for a moment, then whisked the copy of Manchete from my trembling hands. She, in what was a flash of insight, declared that it was not English! She mumbled something else that sounded like English, but held no meaning for me. I gathered by social cues and innuendoes that what I wanted was impossible to ask of her or her co-workers. In frustrated rage and righteous indignation, I stormed the director's office. I explained my communications breakdown with his people and their apparent lack of cooperation. His reply was sobering. None of his people had had any direct experience or contact with Humanists. They commonly worked in mathematic, scientific, business, and administrative applications. He returned with me to the key-puncher as interpreter. He related to me that she wanted my Portuguese samples put onto Standard

IBM code sheets. This meant printing out by hand the Portuguese with appropriate IBM symbols corresponding to the diacritical marks.

It is at this point that I should explain that I was teaching a full course load, engaged in other research, and working alone on the word study. As a result, several months passed from initial contacts to my first batch of cards and "functioning" program in SNOBOL. Actually by late May, my key-punching friend had relented somewhat by allowing me to bring in photocopied material with the symbols penciled in, in lieu of the time consuming code sheets. This was a break through on this front! My program, when run for de-bugging--a process used to check for time and money consuming mistakes--cost me a total of $35 for a 3 minute run. It clearly indicated that one page of approximately 350 words by José Lins do Rêgo was more difficult for the computer to read than a 10 line magazine ad. I happily departed for Brazil, convinced that the computer had no value for me and no place in the Humanities.

This of course is not the end of my story. At best it is a prologue which hopefully conveys, in abbreviated form, some of the tribulations encountered when one first enters this strange world without prior knowledge of the computer

and its people.

It was shortly after my return to campus in January of this year that I met the Computer Center director on my way to the Library. He asked about the progress of my work. I told him bluntly that there had been none and I was thoroughly disillusioned by his wonder machine. I cooly explained that at my previous rate of progress I could obtain a representative count faster and cheaper with the aid of student assistants working a la Keniston. My statement amounted roughly to the equivalent of the chivalric throwing down of the gauntlet. He countered with the point that students would make mistakes, be slower, and more expensive in the long run. He said that he would have an efficient program for me in a day. In my presence on the phone, he described my simple needs to a competent PL/1 programmer who, as if by magic, produced my first successful program within the 24 hours.

This program lists in alphabetical order the words followed by a variable number of columns. The first indicates total occurrences per word, the succeeding ones list the corresponding category in which that word occurs. At the bottom of the alphabetical list there appears the total of all words counted and the total in each category. Pre-

sently I have six broad categories. Category one represents prose fiction and the literary essay. It has ten different samples of approximately 2500 words each. These samples range from regionalist authors like Guimarães Rosa and José Lins do Rêgo to more recent prose writers like Antônio Callado and Adonias Filho. Two covers mass media publications such as O Globo, Curzeiro, Manchete, and other reviews. Three includes material taken from art and the social sciences. Here I have selected items likely to be consulted by History, Anthropology, or Political Science researchers and students. Samples include studies on racism, economic inflation, social development, and revolution. There are at this time ten different samples of 2000 to 3000 words each. Category four contains items from science and technology. There are at present nine samples including a book dear to me, A NOBRE ARTE DE COMER. I wish to add that finding sufficient samples of strictly technical material originally written in Portuguese has not been easy. Five is an aggregate of miscellany, containing such things as program notes, correspondence, signs, folders, and the like. Six deals with samples of verse poetry. I eventually will have a category devoted to theatre. The program is flexible enough to allow the addition of other categories when

needed. There is also another related program which gives the same listing of frequency and range, but only for the samples in each category described above.

All the samples are taken exclusively from Brazilian sources, published between 1946 and the present. Eighty-two per cent of the present samples are taken from the last four years (1965 to 1969). In further additions I intend to maintain the emphasis on the recent years. The samples are selected randomly from a book or periodical. Usually two pages are taken from the front, two from the middle, and two from the last section of the book, but on a few occasions there are samples made up of 2000 running words of text. This is largely determined by the book or periodical's format and the key-puncher's stipulations. After they are punched from the book or review they are returned to me with a print out sheet of exactly what the cards have on them. These sheets are proof read three times for punching errors. The cards containing errors are corrected by pencil, returned to the key puncher, repunched and finally incorporated into the deck. All punched cards are then identified exactly as to the category to which the sample belongs, the code for the author or title of the individual sample, and a sequence number which, in the event of the deck being

dropped, permits easy re-ordering. This sequence number generally corresponds to the line number on the page of the original text. Any given card can be checked against the original text since there is a bibliographical list with the matching identification codes and page references.

My program can punch a card automatically with the Portuguese word on it and the numerical information of frequency and range. With these cards a list can be created of descending or ascending frequency, either numerically or alphabetically. That is to say that under letter A the most frequent words with that letter will be listed in descending order to the single occurrences, then B and C and so on. Actually the data can be manipulated in many ways, but only the above interest me. My word list is exactly that and homographs will not be distinquished. If I wished to know whether conhecido is used as noun or verb, or falarem is future subjunctive or personal infinitive, it could be done by having the machine print out the word in context. This process is very much what a concordance program does and more will be said of this shortly. By not reducing the verb forms to their corresponding infinitives I save money and the effort of teaching the machine how to conjugate, but I believe there may be a pedagogical value in seeing

which forms of the verb are most frequent. In an intensive language course this could possibly expedite the learning and the short range success.

As an aside, the data which I have punched can be searched for idioms--their frequency and range--by preparing a dictionary "look-up." Certain words or phrases are put into a list. The computer is then instructed to scan all the data looking only for those words and expressions, count and print them, indicating their source. This can be executed rapidly, but I must say that it requires a great deal of time on the part of the researcher and programmer. I would like to emphasize that once data is punched, there are innumerable things you can do with it.

Now that you have heard some of my own experiences and a description of my work, I should like to discuss the possibilities for other types of work that the computer can do for us in Portuguese. First I would like to stress the fact that at this time, to my knowledge, there are certainly no more than ten of us in the United States involved in Portuguese and the Computer. There is room and a need for more work in this area. Second I have prepared a bibliography of only a fraction of the work done in other natural languages which could be carried over to Portuguese. There

also exist several publications devoted to computer aided language and literature studies: COMPUTERS AND THE HUMANITIES, HEPHAISTOS, COMPUTER STUDIES IN THE HUMANITIES AND VERBAL BEHAVIOUR, and EDUCOM. It might be good to point out here that the computer used as a tool in language study is not a thing of the future. It is here now and growing very rapidly. It must be remembered, however, the computer is not the answer to all of our research problems. There are very many definite limitations, which, at least this year and probably in the next few years, will not be overcome easily, quickly, and accurately. First among them is easy access to a computer. This of course is changing by the introduction of the time-sharing computer and cheaper rental costs on smaller computer equipment. Second is converting natural language text into machine readable form. Presently the text must be punched into cards and then onto tape or a disk for more compact storage. The development of an effective scanner is what our field most needs at this time. The scanner would be able to read a page of a book or periodical and transfer it directly into the machine thus relieving us of prcofreading--a major source for mistakes. This development as you can well imagine would revolutionize the number of applications and types

of studies available to us. Third is the cost factor which really is a part and parcel of the other items just mentioned. Computers are not cheap. Our model 75 costs $333.30 per hour, however a very large program might only run 30 minutes. Some schools, because of arrangements with the state departments of education are subsidized in teaching computing. Some pay only a small amount, others pay for no actual machine time--only for programmer's services and key-punching.

Returning to the applications of the computer in our field, I would signal out seven main areas that have been tried, tested and approved, so to speák. First in utility, in my own mind, is the word list. The word list can be limited to the preparation of a glossary of repeated and important items in a book, story, poem or whatever for quick and effective adaptation to the class room. Second is the bibliographical list. One prominent use of this is the annual listing of dissertation abstracts prepared at Ann Arbor. Nevertheless, lists of authors, texts, dates, etc. can be prepared and then culled for whatever data is needed. Third is the concordance. This is, in essence, an extended word list. The word is printed out in context with work, page, and line identification. Normally in a

concordance a list of words is assembled which, although counted, does not give context. These are usually common articles and conjunctions which tell us relatively little about the author's vocabulary. A fourth area is linguistic or stylistic analysis. One might search a given writer for clues to influences or literary movements; or one might make a statistical study of the use of the personal infinitive in modern Portuguese. Five is content analysis. This actually is much more used by political scientists than language and literature people. One can prepare a dictionary of basic words and words closely related to them. This is known as a dictionary. The machine readable text is then read and where those words appear, they are printed out in context. A study on the sentiment of death in Huckleberry Finn was prepared in this way. Six is the attribution study. The Federalist Papers were examined by computer and because of the particular use of punctuation and conjunctions the researchers were able to identify those portions written by Hamilton, those by Jay and those by Madison. Seven represents an area well worth our time. It is the preparation by computer of a critical edition. A definitive text with principle variants can be done by collating the material automatically. Finally

there is the experimental realm of CAI. CAI means computer assisted instruction. Annapolis and West Point are working on the use of the on-line computer in the teaching of French, German, and Russian on an experimental basis. This is an area that is expected to develop and it would not be premature to investigate its possibilities at this time.

If you are sufficiently interested in investigating computer applications to your own teaching and research needs, there are at least four publications you could first consult: HEPHAISTOS, COMPUTERS AND THE HUMANITIES, COMPUTER STUDIES IN THE HUMANITIES, and EDUCOM. COMPUTERS AND THE HUMANITIES is particularly valuable for its annual listings of scholars active. EDUCOM is especially good for information concerning CAI.

In closing I would like to cite Louis Milic of Teachers College at Columbia University. He writes:

> It does not seem visionary to state now that within a decade graduate students in literature will be in daily touch with computers and related equipment and that within two decades these machines will be used as routinely in the humanities as the typewriter and Xerox copier. It may be reasonable, therefore, to suggest that an acquaintance with computers ought to be part of the equipment of any person with a liberal education.

I believe, given our present pace, that five years seems

more likely than the forecasted decade by Milic. I believe that we as humanists and teachers cannot afford to neglect the computer and its ever increasing role in education.

John R. Kelly
University of California,
Santa Barbara