# Foreword: Perspectives on the Digital Humanities

**Roberto A. Busa**

During World War II, between 1941 and 1946, I began to look for machines for the automation of the linguistic analysis of written texts. I found them, in 1949, at IBM in New York City. Today, as an aged patriarch (born in 1913) I am full of amazement at the developments since then; they are enormously greater and better than I could then imagine. *Digitus Dei est hic!* The finger of God is here!

I consider it a great honor to be asked to write the Foreword to this fine book. It continues, underlines, completes, and recalls the previous *Survey* of Antonio Zampolli (produced in Pisa, 1997), who died before his time on August 22, 2003. In fact, the book gives a panoramic vision of the *status artis.* It is just like a satellite map of the points to which the wind of the ingenuity of the sons of God moves and develops the contents of computational linguistics, i.e., the computer in the humanities.

Humanities computing is precisely the automation of every possible analysis of human expression (therefore, it is exquisitely a "humanistic" activity), in the widest sense of the word, from music to the theater, from design and painting to phonetics, but whose nucleus remains the discourse of written texts.

In the course of the past sixty years I have added to the teaching of scholastic philosophy, the processing of more than 22 million words in 23 languages and 9 alphabets, registering and classifying them with my teams of assistants. Half of those words, the main work, are in Latin. I will summarize the three different perspectives that I have seen and experienced in these sixty years.

## Technological "Miniaturization"

According to the perspective of technological miniaturization, the first perspective I will treat, the *Index Thomisticus* went through three phases. The first one lasted less than 10 years. I began, in 1949, with only electro-countable machines with punched cards. My goal was to have a file of 13 million of these cards, one for each word, with a context of 12 lines stamped on the back. The file would have been 90 meters long, 1.20 m in height, 1 m in depth, and would have weighed 500 tonnes.

In His mercy, around 1955, God led men to invent magnetic tapes. The first were the steel ones by Remington, closely followed by the plastic ones of IBM. Until 1980, I was working on 1,800 tapes, each one 2,400 feet long, and their combined length was 1,500 km, the distance from Paris to Lisbon, or from Milan to Palermo. I used all the generations of the dinosaur computers of IBM at that time. I finished in 1980 (before personal computers came in) with 20 final and conclusive tapes, and with these and the automatic photocompositor of IBM, I prepared for offset the 20 million lines which filled the 65,000 pages of the 56 volumes in encyclopedia format which make up the *Index Thomisticus* on paper.

The third phase began in 1987 with the preparations to transfer the data onto CD-ROM. The first edition came out in 1992, and now we are on the threshold of the third. The work now consists of 1.36 GB of data, compressed with the Huffman method, on one single disk.

## Textual Informatics

The second perspective is textual informatics, and it has branched into three different currents. Today the two greater and richer ones must be clearly distinguished from the third, the smallest and poorest. I must say that many people still do not realize this.

I call the first current "documentaristic" or "documentary", in memory of the American

Documentation Society, and of the Deutsche Gesellschaft für Dokumentation in the 1950s. It includes databanks, the Internet, and the World Wide Web, which today are the infrastructures of telecommunications and are in continuous ferment. The second current I call "editorial." This is represented by CDs and their successors, including the multimedia ones, a new form of reproduction of a book, with audio-visual additions. Both these, albeit in different ways, provide for the multiplication, distribution, and swift traceability of both information and of a text. Both are recognizable by the fact that they substantially transfer and present, on an electronic support system, words and punctuation plus some operative commands. Because they provide a service and so have a quick return on investment, both have grown abundantly – the first, despite significant obstacles, more solidly and with fewer disappointments than the second.

I call the third current "hermeneutic" or interpretative, that informatics most associated with linguistic analysis and which I would describe as follows. In the electronic *Index Thomisticus* each of the 11 million words is encapsulated in a record of 152 bytes. Some 22 are reserved for the word, and 130 contain 300 alternating "internal hypertexts", which specify the values within the levels of the morphology.

At the moment, I am trying to get another project under way, which will obviously be posthumous, the first steps of which will consist in adding to the morphological encoding of each single separate word of the Thomistic lexicon (in all there are 150,000, including all the particles, such as *et, non*, etc.), the codes that express its syntax (i.e., its direct elementary syntactic correlations) within each single phrase in which it occurs. This project is called *Lessico Tomistico Biculturale (LTB).* Only a computer census of the syntactic correlations can document what concepts the author wanted to express with that word. Of a list of syntactic correlations, the "conceptual" translation can thus be given in modern languages. I have already published, mainly in the series of the Lessico Intellectuale Europeo (directed by T. Gregory of the University of Rome), the results of such syntactical analysis of a dozen words in their more than 500,000 context lines. To give one example, in the mind of St Thomas *ratio seminalis* meant then what today we call *genetic programme.* Obviously, St Thomas did not know of either DNA or genes, because at the time microscopes did not exist, but he had well understood that something had to perform their functions.

## Hermeneutic Informatics

This third sort of informatics was the first to come into being, with the *Index Thomisticus* project, in 1949. It brought the following facts to my attention. First, everyone knows how to use his own mother tongue, but no one can know "how", i.e., no one can explain the rules and no one can list all the words of the lexicon that he uses (the active lexicon) nor of that which he understands but never uses (the passive lexicon).

What scholar could answer the following questions? How many verbs does he know at least passively? Which and how many of them are always and only transitive? Which and how many of them are always and only intransitive? Which and how many of them are sometimes the one, and sometimes the other, and what is the percentage of each? Lastly, which contextual situations characteristically mark the transitive or intransitive use of the latter?

Second, there is still no scientific grammar of any language that gives, in a systematized form, all the information necessary to program a computer for operations of artificial intelligence that may be currently used on vast quantities of natural texts, at least, e.g., for indexing the key words under which to archive or summarize these texts to achieve *"automatic indexing – automatic abstracting."*

Third, it is thus necessary for the use of informatics to reformulate the traditional morphology, syntax, and lexicon of every language. In fact all grammars have been formed over the centuries by nothing more than sampling. They are not to be revolutionized, abandoned, or destroyed, but subjected to a re-elaboration that is progressive in extent and depth.

Schematically, this implies that, with integral censuses of a great mass of natural texts in every language, in synchrony with the discovered data, methods of observation used in the natural sciences should be applied together with the apparatus of the exact and statistical sciences, so as to extract categories and types and, thus, to organize texts in a general lexicological system, each and all with their probability index, whether great or small.

Hermeneutic informatics hinges on the Alpac Report (Washington, DC, 1966) and, now, this perspective is perhaps awaiting its own globalization. I have already said that hermeneutic informatics was the first to come into existence. Shortly afterwards, in the early 1950s, if I am correct, the move toward automatic translation started. The magazine *MT – Mechanical Translation* was started at MIT, launched, I think, by Professor Billy Locke and others. The Pentagon financed various centers. I was involved in this. I connected the Anglo-Russian project of Leon Dostert and Peter Toma (of Georgetown University, Washington, DC) with the Computing Center of the Euratom of Ispra, which is on Lake Maggiore in Lombardy. My contributions were on an exchange basis. I supplied them, from my laboratory at Gallarate, with Russian abstracts of biochemistry and biophysics in Cyrillic script on punched cards, a million words. These were translated with the Georgetown programs. The translation was sufficient for an expert on the subject to be able to evaluate the merits of a more accurate translation done by hand, i.e., by a person's brain.

Unfortunately, in 1966, as a result of the Alpac Report, the Pentagon cut off all funding. This was not because computers at that time did not have sufficient memory capability or speed of access, but precisely because the information on the categories and their linguistic correspondences furnished by the various branches of philology were not sufficient for the purpose. The "machine" required greater depth and more complex information about our ways of thinking and modes of expression!

## Future Perspectives

In certain respects this was a boon. In fact, as this volume, too, documents, the number and devotion of those few volunteers, who in almost every part of the world have a passion for computational informatics, increased. They are not an organized army, but hunters who range freely, and this produces some obvious side effects.

It also provokes a valuable consideration. Namely, it makes us realize that no single research center ever seems to have been able to answer, alone and completely, the linguistic challenge of globalized telematics. It seems that the answer to globalization, at least in principle, should be global as well, i.e., collective, or rather undertaken by one or more supranational organizations, this for the obvious reason of the commitments required. Speaking is an interface between infinity and the cosmos, between time and eternity, and is evidence of the thirst for knowledge, understanding, possession, and manipulation of everything, according to one's own personal freedom, but on track with a common ballast of logic and beauty. Speaking must thus be taken seriously; it is sacred, as is every human person. We are far from having exhausted the precept inscribed on Apollo's temple at Delphi, "Know thyself." It seems, therefore, that the problem must be attacked: in its totality – with comprehensive, i.e., global, research; collectively – by exploiting informatics with its enormous intrinsic possibilities, and not by rushing, just to save a few hours, into doing the same things which had been done before, more or less in the same way as they were done before.

It seems that the attack on the Twin Towers of New York City on September 11, 2001, has brought in an unforeseen season *of lean kine.* In Italy, as everywhere else, this has meant reductions in public funds for research. This period will pass, as surely as all the others we have experienced. Such reductions were also in evidence at the Exploratory Workshop on Computer Texts, which the European Science Foundation of the European Union held at Strasbourg on June 14 and 15, 2002. On the one hand, these cutbacks in finance are certainly worrying for the many operators of computational linguistics, which today is fragmented, but on the other hand, it could also lead to the according of priority to a definitive solution of the linguistic problem, like that which could facilitate the fulfillment of the globalization of economic exchange.

## A Proposal

I should like to summarize the formula of a global solution to the linguistic challenge that I presented at the above-mentioned conference at Strasburg, much as if it were my spiritual testament, although I am uncertain whether to call it prophecy or Utopia.

I suggest that – care of, for example, the European Union – for every principal language A, B, C, D, etc., from each of the principal university textbooks in present use in the various disciplines (for these represent the present state of what is knowable) there should be extracted its integral "lexicological system" (with the help of the instruments tested in the *Index Thomisticus).* In it, two "hemispheres" of each lexicon should be distinguished, the few words of very high frequency present in every argument which express the logic, and which are sometimes called "grammatical", and the very many words of relatively minor frequency that specify messages and arguments. The systems of each text of each discipline in each language should be integrally compared with each other by isolating and specifying both the coincidences and divergences with their quantities and percentages, including those of hapax, extracting from them, i.e., mixing them, in one single system of the same language, a system which totals statistically and with percentages both how much they have in common and their co-respective divergences.

Then these statistical summaries of the various languages A, B, C, D, etc., should be compared with each other, with the same method of reporting their convergences and divergences, with quantity and percentage in a single system. (One has only to think of how much data could be published *a latere* as valid and useful documents, although to be constantly updated, for example for contrastive grammars, etc.)

Thus there would be on the computer a common interlingual system consisting solely of strings of bits and bytes with correspondence links both between convergences and divergences in themselves and between each other. It would be a sort of universal language, in binary alphabet, "antiBabel", still in virtual reality. From this, going in the reverse direction, there could be extracted in the respective alphabets of the individual languages A, B, C, D, etc., the number of words and expressions, situations of morphology and syntax, of each language which have found correspondence in other languages, and the number of those which have not. The number of such correspondences thus extracted (lexicon and grammar) would be a set of "disciplined" basic languages, to be adopted for the telematic use of the computer, to be also printed and then updated according to experience.

In input, therefore, everybody could use their own native disciplined language and have the desired translations in output. The addressee could even receive the message both in their own language, in that of the sender, and in others.

In addition, the problems of keys and privacy will have to be solved, as (one step at a time!) will those of the phonetic version of both input and output.

These thoughts have formed gradually in my mind over the years, starting from the realization that my programs for Latin, which I always wanted broken up for monofunctional use, could be applied with the same operative philosophy to more than twenty other languages (all in a phonetic script), even those that do not descend from Latin, such as Arabic and Hebrew, which are written from right to left. I had only to transfer elements from one table to another, changing the length of fields, or adding a field. (However, I cannot say anything about languages written in ideograms or pictograms.)

## Conclusion

In conclusion, I will therefore summarize the third perspective, that of textual hermeneutic informatics, as follows. The first period began with my *Index Thomisticus* and ended, though not for me, with the Alpac Report. The second, after the Alpac Report, is the parcelization in progress of free research. The third would begin if and when comparative global informatics begins in the principal languages, more or less in the sense I have tried to sketch here. Those who live long enough will see whether these roses (and thorns), which today are merely thoughts, will come to flower, and so will be able to tell whether they were prophecy or dream.