Informatics and New Philology

Roberto Busa SJ

Aloisianum, 21013 Gallarate, Italy

Abstract: This paper considers the impact of informatics and the new technology on the field of philology. A brief overview of developments in the field is given, and the present bottleneck of computational linguistics are considered. Also discussed are some of the challenges posed by natural language processing, and the ways in which we can rise to the challenge.

Key Words: informatics, computer science, philology, lexicography, linguistics.

I.

Let me say at the beginning that the title “Informatics and New Philology” should not be understood as implying that traditional philology is to be forgotten, nor as suggesting that we need to start again from scratch. That is certainly not the case.

The science of philology was not born with the computer. The new philology will be a continuation from its predecessors, but with a quality-leap and new dimensions. If so much was done without the computer, how much more should the young people of today be able to achieve with the computer? Such is the meaning of the title of this paper.

Computational linguistics wards off that threat to our writing skills caused by the growth in audio-visuals. The computer is in fact an expression of mankind, no less expressive than his speech. The philology which works with these new tools, using this new electronic alphabet, works for its own growth. Our generation has not done everything: for the young people there are still immense open spaces. This is what I would like to show.

II.

First of all, I would like to summarize the developments in philological methods that the computer has been responsible for over the past forty years — that is, from 1949 when I first began to grapple with those gadgets!

Working on physical entities which the philologist uses and classifies, the computer has served above all to demarcate the borders between the graphic and the semantic. Furthermore it allows one to recheck and redocument every conclusive affirmation whenever the need arises.

What has perhaps not yet sunk in, is the huge change in dimensions that has occurred: the computer enables us to perform a complete census on texts of tens and hundreds of millions of words. For so long the philologist, being able to work only with pen and paper, had to be happy with a mere sampling, like a bee flitting from flower to flower. We all know how unreliable sampling becomes as soon as semantics is called upon.

The reason for another great quality-leap lies in the fact that, starting from the census-analysis of graphic signs, the computer has imposed the systematizing of words at the level of forms, before and instead of simply gathering them together in the lexical unities we call lemmas. This recognition of the distinction, opposition and composition of these lexical levels — forms and lemmas — opens totally new possibilities for philology: it is true that this distinction had always been present in writing grammars but not in the general or authorial lexicons.

Roberto Busa, Jesuit, born 1913, is a retired full professor of philosophy. His most important research area is linguistic analysis. Among his publications are Index Thomisticus, 49 vols., Stuttgart, 1974—79; and Fondamenti di informatica linguistica, Milano, 1987.

It follows from the above that today the technique of lemmatizing can be more conscious and systematic. It makes us aware of how many zones there are, and of the generative mental language which we always use and never explore. It also brings us to realize the number of speech zones which we still need to plan and map.

I would say the same for the complex and exquisitely lexicological phenomenon which we call homography. It deals with words in isolation, in their generative potentiality. The number of words with identical spellings but with different meanings is greater than one would perhaps expect. Their classification depends on the interplay of various elements: if the word is a form or a lemma; grades and levels of difference in meaning; the diverse sources of identical spellings; the interconnection between homography and homophony.

Finally, I would emphasize the demarcation of the morphological from the syntactical, complicated as it is by the features we have just mentioned. Every word when isolated, that is before it is used in a sentence, is already a package, a system of meanings. Many more layers of meaning are added to it by the syntax within the sentence. It is still not generally recognized that the *nomen* and *verbum* are morphological categories whereas the *nomen* — *substantivum* and the *nomen* — *adjectivum* are basically syntactical categories, as are all other parts of speech.

III.
The computer reveals its limits exactly where philology is at its strongest: the interpretation of the whole from the whole. This consists of the intuitive insight into the formula, form, value and meaning of the whole. From this understanding of the complete phrase we work down to specify the values of the words and clauses which make up the phrase. In some real sense here, it is the synthesis which precedes the analysis.

With the computer, however, we begin with a piecemeal analysis of the groups of bytes with which we have formalized the graphemes of the written text. In the computer the strings of bytes are and remain an aggregate: the only link they have is that of their respective positions. With the computer we need to add codes or tables in order to correlate one group of bytes with another, due to the semanticity of the words rather than just the position of the bytes. It is merely a convention that in our computers a certain string of bytes expresses a certain information-content. The global meaning of the whole is represented to us by the ensemble of its words and sentences.

In fact, in every whole which is a “system” and not a mere heap or mass, there is, superimposed on the components, a specifically chosen distribution, arrangement and organization. Such an organization may be called design, pattern, formula, system, orchestration, engineering, form and idea. When building a house we can buy all our materials, but we must still pay an architect who tells us how to put them together. Such is the case with inventors, composers, sculptors and programmers, etc.

However, a computer program would be able to derive the values of the words from the syntax of the whole only if this global syntax has already been formalized and loaded into the program, either because in the text some elements have been found capable of signifying it, or because of the appropriate codes added by human intervention. I know of no other way a program could simulate the intuitive interpretation of the whole. It is already easy and common to formalize in computer tables the categories that can be recognized from the morphology, i.e. the graphic structure of each individual word. However, no computer table can provide the variable categories added to each word by the various sentences, unless the program contains algorithms corresponding to the global meanings of each sentence, or unless it contains the much more complicated algorithms which enable the computer to grasp and formalize the global meaning of the whole which includes all global meanings of all sentences.

Syntactic analysis, formalized grammar, automatic lemmatization, and artificial intelligence are slowly confronting the problem of semantic processing. This problem challenges linguistics more than informatics: it depends more on our knowledge of language than on the techniques of programming. Its solution would represent an enormous development not only in informatics but also in linguistics. It would ultimately lead to a mathematical model of human expression.