Half a century of literary computing: towards a "new" philology

Busa, Roberto

Historical Social Research, Vol. 17 — 1992 — No. 2, 124-133

HUMANITIES COMPUTING

Half a Century of Literary Computing: Towards a »New« Philology

*Roberto Busa SJ (Gallarate/ Italien)**

Summary

§§ 1-6 Introductory remarks

I. In what sense can computer do so little?
§§ 7-11 In processing text as text, because our linguistic information is inadequate

II. Why our philology, up to today, is inadequate to substantiate artificial intelligence in text processing?
§ 12 Incidental factors
§§ 13-16 Inner reason: for programming a computer our mind needs to analyze micrologically its macro-intuitions.
§ 17 Such reflexive introspection can be done scientifically only by a larger and deeper inductive analysis of natural texts.

III. The »new« philology
§§ 18-19 It must be able to formalize the global meaning of a textual set.
§ 20 Some already conquered new philological data.
§§ 21-22 Trends of research: in sentence types, styles, statistics ..,

IV. Conclusion §§ 23-24

I feel proud having been invited by Prof. Ott to talk to your Institute, as your activity is universally considered as solid, efficient and humble, meaning by this last word that your production is larger than its publicity.

I feel grateful to God, as 30 years ago, precisely on the same day as today, my initiative, already 14 years old, was »confirmed« at the »Kolloquium über Maschinelle Methoden der Literarischen Analyse und der Lexikographie«, organized here in Tübingen by Prof. W. Schadewaldt, IBM Deutschland and me. Previously I had reported about it in Bad Nauheim, Oct. 1956 at the annual convention of the Deutsche Gesellschaft für Dokumentation.

* Protokoll des 50. Kolloquiums über die Anwendung der EDV in den Geisteswissenschaften an der Universität Tübingen am 24. November 1990.

1. What I will now say is a cocktail of status artis and testament, as I am 77 years old. Moreover it is a summary of what I have experienced, and not of what has been written by others.

In fact I started to explore how to automatize linguistic analysis in 1946. I started to play with IBM punched cards machines in 1949. I have punched and processed 6 million cards. I started to use IBM computers as soon as they existed. I have put in I/O more than half a billion records, containing either one line or a word with its »internal« hypertext: of texts of 18 languages and 8 alphabets. I have photocomposed by computer 80,000 pages. I have entered into computer by optical scanner 12 million characters. Finally I have compressed, without any loss of information, into 120 million bytes on a CD-ROM the 1,630 million bytes of the Thomistic Latin corpus with its hypertext. I also founded, 15 years ago, a school of informatics for humanity students at the Catholic University of Milan, the GIRCSE: »gruppo interdisciplinare ricerche computerizzazione segni espressione«.

2. About text processing I shall not report either on the hardware or on the software side, but only on the side of the philological analysis of the texts to be processed.

The syntagma »computer and the humanities« refers not only to texts, but also to speech, fine arts, music, mime, theatre, film making etc.: but for the sake of brevity, I shall talk only of texts, tacitly and analoguously implying all the other realms. I have calculated the risk of being qualified as simplistic. I am sure I am not.

3. I am using the word »philology« meaning by it all sciences defining how we speak and how we write.

I use the syntagma »linguistic analysis« attaching to »analysis« its generical value and none of the recent historical specific values. I call »analysis« any census of a text, aimed at detecting and classifying the elements and structures and categories existing in a text.

Deliberately I try to avoid all terminologies specifically adopted by philosophers of language: I do not want to portray a language, but just to draw its lexical map.

4. Thinking about how we do think, speaking about how we do speak, writing about how we do write, is an introspective, reflexive, inner, interior activity: the old classic »via interioritatis«.

5. In the 1950's newspapers were contrasting the rude and crude technology to the gentle and frail humanity: as if machine could endanger human thinking. Today specialization lead to a subtler and deeper problem, that of the incommunicability between disciplines, a kind of entropy and decadence of the culture. It invades ecclesiastical sciences too. Splitting the

knowledge of man into isolated parts, implies breaking the unity of man and of men.

Sciences, humanities, technologies, business, politics, religious meditation must be composed, not opposed. Isn't there any human field which is neither spoken of nor written about? All technology is human expression no less than poetry, just differently aimed. Cutting off all continuing between human activities, e. g. religion from science, business from ethics, is anti-scientific: it hides some irrational steps, like trying to break one operating system (e. g. a running engine) into two or more machines. Being a priest, people often consider my presence in computer science as exotic, like if you met a camel in your Marktplatz. But it is precisely as a priest that I am doing what I do. In fact analyzing texts leads to realizing the presence of the mystery of God at the roots of human understanding and talking. Moses is on my side: Deuteronomy 30, 11-14 »The word [inviting to chose life and good or death and evil] is very near to you; it is in your mouth and your heart ...«. When, July 18, 1956, Norbert Wiener visited me at Gallarate, we agreed about it. Later he published a booklet »God and Golem - Cybernetics impinges on religion« (MIT 1964).

6. I have divided this speech into three chapters: I. in processing human discourse, computer can still do very little II. why? III. because a new philology is needed.

Frankly, I feel sure of what I am saying, but, as every year I am more aware of how difficult it is to evaluate and administer our personal certitudes, I ask you kindly to listen to me critically.

I. In what sense can computer still do so little?

7. Computer services are certainly so monumental as to characterize our era, whenever processing digits and figures for all sciences and for any kind of administration and business. In processing letters and any other kind of sign beside digits, i. e. in processing texts, computers have also started to perform current services in three fields:

- copying and transmitting: e. g. photocomposition and telefax;
- office automation, e. g. spellcheckers and desktop publishing;
- information retrieval, i. e. in consulting data bases.

All those have two features in common: a) computer processes characters purely as graphic signs b) computer has been equipped to process some semantic contents of single words or strings of few words: e. g. lemmatizing and parsing.

8. But in processing a text as text, computer science is still detained and entangled in painstaking laboratorial research. I call »text« any »discourse« e. g. an essay, an article, a novel, a patent and even a computer program ...

I do not consider as »text« the listings of items, a telephone directory, a train time-table ...

A text is a »system«: i.e. an assembling - of different components - adapted and connected to each other - to constitute a unit - aimed at operating a definite performance.

In all text there is a global unitarian meaning, i. e. a one global, defined definite and definable, architecture of thoughts, resulting from the assembling of as many sub-units as there are parts, chapters, paragraphs, sentences, clauses.

9. As a proof that computer science has not yet subdued texts as texts, I take the fact that there are not any practical large scale applications of fully automatic indexing and of fully automatic abstracting. »Language industry« will explode the day in which man is able to write programs ready to perform it currently. At present, that is still detained in laborious and laboratorial attempts. Even the history of the »machine translation« confirmes the same, since when the ALP AC Report (Washington DC 1966) stated that the previous audacious and impetuous attacks on it were naive. It is not the computer which is lacking - memory capacities and programming skills have already been many years more than enough - but philology. The linguistic data we are feeding into computer for text analysis are not yet adequate to the computer potentials and ways of operating. I remember two writings of mine: »The impact of Cybernetics on the Humanities« (Proceedings of the Jurema 1972 - Int. Symposion) and »Why can a computer do so little?« (ALLC Bulletin Vol. 4, 1976, pp. 1-3). The latter was the results of my meetings with the Bundes-Archiv in San Francisco XIV. Int. Congress of Hist. Sciences, August 1975 and later in Koblenz. The former I had read in Zagreb, April 1972.

10. In one word: our philological knowledge is not yet ready to formalize the global meaning of text units and sub-units. Ill explain it in the next chapter.

11. I conclude this chapter in the following way:

- computer science as far as artificial intelligence on texts, cannot go further without an enhancement and deepening of our philology;
- no present grammars, no present vocabularies provide enough information for programming practical services in automatic processing of texts;
- even less adequate to it is the knowledge which each educated programmer has, or is able to derive from the grammars and dictionaries, of his own language, today.

11. Why our philology, up to today, is inadequate to substantiate artificial intelligence in text processing?

12. First of all there are some external and incidental factors.

- Computer has been nursed by English language, the structures of which are simpler than those of many other languages.
- Too many programmers don't realize how big a mistake it is to process words as they process digits: when their semantics is processed too, words are deeply heterogeneous even within the same sentence, while within the same file digits and numbers are homogeneous.
- Two metaphors explain two more defects. First, too many researchers pile up one km of algorithms on a base of one cm, instead of building up at first one cm of algorithms on a base of one km, and then another cm on top the first one over the entire km and so on ...
- Second: one hundred contractors are asked to construct a road in a jungle. Each one constructs its first km, but no one constructs the second km, no one the third km and so on ...
- Due to some academic urgency for publishing much and quickly, too many investigations seem to be just miniature models of research works which would be wonderful if executed in their natural dimension ...
- In all administration, computer lessens human labour because a program, once written, can be repeatedly used. However it would be an error to imagine that in research computer may lessen the quantity of the personal work and effort. Relieving the researcher from the tasks of copying, searching, grouping and counting, computer demands that he examines and checks giant quantities of output and condenses organizing and classifying operations in a short time: i. e. computer allows the researcher to perform much less secretarial work, but imposing on him higher quality decisional intelligent activities in closer time. In the end, using computer, a researcher has to work much more than before ... For the Index Thomisticus, we used no more than 10,000 machine hours (including punched cards machines), but 1,800,000 man hours.
- Finally, another unawareness - largely diffused, I am afraid - is also responsible for stagnations in computerized linguistics: that too often computer is used to reach the same targets as before, using the same methods as before ... Philologists must create new strategies for new goals, when using computers. The skills of a taxi driver are of no use for piloting a jet plane and traffic regulations in the skies are tremendously different from those on ground.

13. I shall now hint at the basic inner reasons. Discourse and text are one extremity of the operational arc thinking - speaking. The two extremities are heteroclitous and irreversible. They are one of the realities verifying the mysterious opposition of the inseparable couple active - passive, au-

thor - work, generating - generated: see the copyrights, the royalties, the patents ... Knowing is much more than simply memorized information. It is also unrestrained curiosity, drafting people to safaris in the jungles of the unknown: but I should say that for our thinking there is not the unknown but the not-yet-known.

The »sign« per se does not exist in knowledge but in its communication. Speaking is a system, where words are interfacing. »A« the transmitter, to »B«, the receiver, with »C«, the word, transfers »D«, the concept or message. A, B, C and D are mutually exclusive. But the four altogether are the transparency between A and B.

The graphic and semantic correlation C + D, sign and sense, word and concept, is free, creative, spontaneous, socially conventional and historically evoluting: not commanded by the nature of the signified nor of the signs, but invented and adopted by the social thinking power.

Speech and text are physical entities - strings of signs - linear, closed, fixed, terminated. Thinking - somewhere within our body: who knows where? - is something whose being and acting dimensions are just opposite to the former: generative, multi-dimensional, co-present, having a centre »diffused everywhere«, assaulting not only the present, but also - and even more so - the no-longer and the not-yet...

Signs of it can be seen in all spoken and written sentences. The elementary unit, the molecule, of our expression is sentence, not word: words are like its atoms. Each sentence is a system, an architecture.

In thinking and consequently in language, two functions or levels emerge: one is the power, the vehicle, the conveyor belt, the active, aggressive, dominating, pushing, logical and criticizing power ...; the second are the messages, the contents, the »ideas« ...

The former are certainties »with which« we think and speak i. e. which make us think and speak. The latter are »what we think of and what we speak about«.

This statement implies that there are two types or levels of certainties: the vital or generating ones, and the cultural or generated ones: i. e. the power-author ones and the product ones.

14. Semantics and hermeneutics are reflexive sciences of how we are able to understand i. e. to go and get and receive and grasp the »concepts« or »messages« of another person and consequently the style and mode of his »spirit«, from his words and/or from his works, as in all aesthetic interpretation and criticism of music, fine arts, poetry etc. I would not include in it the so called »de-programming«, i. e. when from the performances of a system one tries to reconstruct its program: in fact, the author-work correlation runs between the programmer and the program, but not between the program and its operations.

15. Artificial intelligence on human texts is nothing other than the hermeneutical process delegated by man's mind to a machine in the form of a program. To write such a program, using his inner understanding power, one must first reflexively understand it, i. e. understand how he understands. Moreover, he has to analyze and break such process into its elementary factors and steps, so that he can »express« them in electronic bits.

Our era has been characterized, since the advent of electronics, by the fact that from the sciences of the »macro« we went to the sciences of the »micro«, having constructed instruments to penetrate into the recesses of the structure of the matter. In addition and consequently, all computer science is nothing other than the mind demanding to itself to anatomize and mathematically formalize its macro-logic intuitions into its elementary micro-logic units and steps. So, the difficulties of artificial intelligence are in the »intelligence« not in the »artificial«: i.e. we do not understand enough how we understand, not enough as to be able to spot our inner logical fibres and steps when understanding ...

Micrological analysis is first of all an exquisitely philological affair.

When we are able to achieve such an analytical, reflexive micro-understanding of how we understand, there will be no problem in expressing it both in words and in machine form.

16. It seems that all our inventing and expressing activities start from a sudden and illuminating intuition of a whole new set of various entities and operations producing a definite result. From such intuition we descend then to determine one by one all its details.

In reading a text of others, we get first the meaning of at least some words here and there, till we suddenly grasp the global meaning of the set, sentence, chapter etc. From such intuition of the global sense of the whole, we then descend to recognize each individual value in and from its context. In both these »virtuous circles« the basic and starting power of mind seems to be that of dominating at once by a glance the unity of the architecture of the set or system. Such intuition seems to be the essential one in the mental process of hermeneutics.

The global unity of the text is, in any case and certainly, expressed, sometimes redundantly, by the text as a whole i. e. by all its words.

But human mind is also capable of summarizing (indexing, abstracting, telegraphing), i. e. of comparing the unity of the set with its single components.

All such operations will be programmable when we achieve their micrological analysis. Even if we do not achieve it or not fully, many good computer services will materialize from having tried it.

17. Science is social. Scientific introspection must also then be social. The only possible way I know for such a paradoxical »social introspection«, to be effective, is what I call linguistic analysis of natural texts.

The »social« always needs the »sign«. And the use of computer not only allows exhaustive censuses, but also gives protection against the dangers occurring when someone builds up philosophies based on personal intuitions only.

Between expressing himself by phonemes and graphemes or by »bits«, the only conceptual differences are that computer - the »stupid« machine by which the intelligent mind extendes itself - demands to work micrologically and commands a rigidly systematic coherence and consistency of everything.

III. The »new« philology

18. I am speaking of a »new« philology, in a evolutionary but not revolutionary sense. I may describe it either as the one which points to the micro-analysis of our macro-cognitive processes, or - and it is the same - the one which points to formalize for computer use the global meaning of textual sets.

The global meaning is not purely the aggregate or sum of the meanings of the individual words composing the text, but the »form« of the distribution superimposed to them. The builder of a house has to pay the suppliers for all materials, and also the architect, for the idea and design of the house, though it is not an additional »material«.

The textual global meaning is already formalized by all the words existings in machine readable form. The problem is how to detect in the text, or to insert into it, a few words or other signs characterizing the global unit of the set.

19. The method cannot but be inductive. We have to renew all our linguistic definition and classification, censusing all linguistic elements, one by one, on many very large and different natural texts. Personally I would start doing it on the parts of speech.

20. I shall now summarize some new acquisitions emerged from my works over half a century.

- A classification of graphic text symbols: letters, pro-letters, in-letters, digits, punctuations, graphotypes, text-typology-codes, operating-codes ...
- Reasons and methods for pre-editing the various text-typologies.
- The concepts and values of the »lexical system* of a text: forms, lemmas, themes.
- The necessary distinction between the morphological word categories (the tripartite division: nominal inflections, verbal inflections, invariable particles) and the syntactical categories: the parts of speech.
- The necessity of lemmatizing and its procedures.

- The census and classifications of homographic words (and of the homophones) as per types, frequency, source, recognition ... e. g. in Thomas Aq. 56.84 % of the words are homographic.
- The census and classifications of inflectional endings of the word forms. E. g. in Latin we counted 3,924 semantically different endings, reducible to only 860 graphically different ones, as many represent more values.
- The census of morphematic segmentation of the lemmas, into initial, middle and closing elements (prefixes, themes, suffixes ...). In Thomas Aq. it came out that all his 9 million words are combination of only 1,882 different groups of letters (plus the 860 endings) occurring at least once every 100,000 words. (There are 2,384 different strings occurring more rarely, in circa 40,000 words).
- The classification and census of »semanticity types« i. e. of the heterogeneity of the words, i. e., of the different relations between word and concept. E. g. in Thomas Aq. 2.6 % of the words are proper names; 1.24 % are deictic; 10.62 % are »vicarious«; 6.67 % signify tangible objects; 0.33 % signify invisible objects; 46.83 % signify aspects and 35.18 % relations.
- The frequency counts of lemmas so different from those of forms. E. g. in Thomas Aq. one fourth of the »common« lemmas occur once or twice; two fourths occur from 3 up to 100 times; one fourth occurs more than 100 times, i. e. 2,263 up to 1,000; 665 from 1,000 up to 10,000; 169 over 10,000, up to the 466,781 occurrences of the verb »sum«.

I consider that these facts are already seeds of chapters of the new philology. They are documented in my publications.

All these investigations have been interactive. But the natural force of things always obliged us to do by hand, or better by mind, the first spotting of types and of boundaries between them. The same force has constrained us to start analyzing the morphology and semantics of single words in isolation, postponing to a necessarily second stage their syntactical characterization.

21. Finally, I shall hint at some lines of research upon which I have stumbled during the above mentioned enquiries, and which appear to be necessary for text processing. Some of them are or have been attempted here and there by others, but, if I am not mistaking, all are still laboratorial and tentative.

- The problem of how to detect automatically, among the graphically co-occurrent words, both the syntagmas (i. e. the multi-word-lexical-units) and all words semantically correlated into clauses within a sentence.
- The problem of processing the tacitly implied words: e. g. »You like it, I don't«.
- The problem of how to single out and process all metaphorical uses of words. I have a booklet printed in Lyon 1533, which published writings

of classic grammarians, listing 217 different metaphors and rethorical figures ...

22. The last paragraph already introduces syntactical analyses. This is a gigantic field, where many spaces are still untouched.

- Automatic parsing, semi-automatic lemmatizing, automatic disambiguation of equivocal words, have already been explored for many years.
- But what about classifying types of sentences and sets of sentences? What about classifying types of »reasonings«?
- In some learned fields there is much speaking about »literary genders«: but what about micrologically (and not only by tasting and testing) defining, listing, classifying the features which differentiate one text-gender from another?
- Even statistical linguistics seems to me a field still raw and unripe. Has someone already enquired how the significance of frequency counts of words is affected by the fact that words are so deeply heterogeneous, and from as many sources as are semanticity types, homography, syntax, metaphor, text gender?
- Furthermore the field of text statistics which points either to its authorship or to its chronology, still needs to be founded, at least by locating and listing those text features which either cannot be found together in the writings of one and same author, or within him are possible in only one definite time-sequence.
- All that implies that we are also still waiting for a mathematical formula of the style. This, at least ideally, cannot but be global. See the booklet I edited »Global linguistic statistical method to locate style identities« - Proceedings of an int. Seminar, (Gallarate June 1971) Rome ed. Ateneo 1982. But aren't there also in a text single »checking points«, as is the pulse for fever, acting as style fingerprints?

As far as I know, we do not yet have a scientifically documented list of such features.

IV. Conclusion

23. In one word, the new philology will explode into a »language industry« when our mind has analyzed micrologically the elements and the steps of the macrological intuitions by which we grasp the global meaning of sets of words composing a text.

New interactive methods and strategies of linguistic research are expected. They will be the spring, the engine, the soul of such new philology. Young people find in it enormous quantities of work to which to apply their creative ingenuity.

24. The following pseudo-syllogism give me hope: Computer is the son of man. Man is the son of God. »Ergo« God is the grandfather of the computer ...