ASSOCIATION FOR LITERARY AND LINGUISTIC COMPUTING

CHAIRMAN: PROFESSOR R. WISBEY

Department of German, King's College, University of London
(Founder and former Director, University of Cambridge
Literary and Linguistic Computing Centre)

The principal purposes of this international association are to further literary and linguistic research by computer and to provide a means of communication for all those concerned with such research.

COMMITTEE

OFFICERS

Chairman	- Professor R. Wisbey
Honorary Secretary	- Mrs J. M. Smith
Honorary Treasurer	- Professor D. A. Wells

MEMBERS

Professor A. Zampolli	- *Italy*
Dr F. de Tollenaere	- *Netherlands*
Professor W. Lenders	- *West Germany*
Mr M. H. T. Alford	- *United Kingdom*
Professor I. T. Piirainen	- *Finland*
Dr W. Martin	- *Belgium*
Dr F. V. Spechtler	- *Austria*
Mr S. Bootle	- *United Kingdom*

SPECIALIST GROUPS - REGIONAL BRANCHES - SYMPOSIA - SUMMER SCHOOLS

REPRESENTATIVES in Austria, Belgium, Canada, Denmark, Finland, France, Germany (West), Hong Kong, Hungary, Iceland, India, Israel, Italy, Mexico, Netherlands, Norway, Sweden, Switzerland, United Kingdom, USA (West)

Enquiries should be addressed to the Honorary Secretary:

Mrs J. M. Smith
6 Sevenoaks Avenue
Heaton Moor
Stockport
Cheshire SK4 4AW
England

GUEST EDITORIAL:

WHY CAN A COMPUTER DO SO LITTLE?[1]

R. Busa, S.J. *(Member of Advisory Board)*

* * * *

Abstract

While in business, technology, and scientific computation, the electronic era is already providing us with monumental services, in processing texts by computer we are still at some laborious and primitive stage. In fact, we can operate on texts only in terms of individual words, which is a rather poor performance. Automatic indexing, automatic evaluating, and automatic summarizing are still to come. The reason is the fact that the levels of meaning which a word has when it is inserted in a sentence are more numerous than those it has when taken alone, isolated from context.

* *

I was the first one, back in 1949, who started to make use of automation for processing non-numerical, literary information, and consequently I have been able to follow its explosion during the last twenty-five years. During that time, as you may know, I have personally processed by computer some two million lines of text, a total of some fifteen million words in nine languages and four alphabets (Latin, Greek, Cyrillic, and Hebrew). Two-thirds of it are the *Index Thomisticus*, a linguistic analysis of the 118 works of St Thomas Aquinas plus sixty-one works of other authors ranging from the ninth century up to the sixteenth century. It is a product of much work of a large team; for five years I had up to sixty-five co-workers. We estimate that it entailed one million man-hours. I have already photocomposed by computer and then printed by offset the first thirty-one volumes of the *Index Thomisticus*, some thirty-six thousand pages of encyclopedia size. I am currently working on the processing and photocomposing by computer of the last twenty to twenty-five volumes. I publish my *Index Thomisticus* as a document providing scholars with facts, primarily for linguistic sciences.

In practice, what we can do today by computer for publishing documents is no more than processing their indexes, manually preparing the input for them. In other words, in processing human information we cannot currently do by computer much more than (a) to transcribe on tape a full text, and (b) to operate on it only in terms of individual words, or, when manually predetermined, individual keywords.[2] Conceptually, this is rather a poor performance, and it becomes even more intriguing when we think of the giant and endless developments of computer uses in business, technology, and scientific computation. I therefore decided to express as clearly as possible some of the implications, i.e. why the use of the computer is still detained at some primitive and laborious stage in processing languages, while its services in other fields are already monumental.

Electronic data processing marked the beginning of a new era in the transfer of human information. A new era, to me, is when some basic social factor makes a qualitative jump and starts to have new dimensions and new operational parameters. At Gutenberg's time typesetting started a new era in the distribution of human knowledge. Today, we have made another jump: we are now able to use an electronic alphabet which can be processed by machine at 'electronic' speeds and distances. But we are still at the starting point of the new era as far as language processing

is concerned. In these new technologies when we consider not their starting points but their trends in research and development, we realize that the computer challenges scholars to new dimensions of scientific research; by this I mean to improve the quality of the research itself. In fact a computer is a machine operating physically on physical entities taken by man as symbols. Consequently, it is just a tool, an off-line continuation of man's fingers, as fingers may be described as the bodily computer on-line to the mind. It is man's ingenuity which has to feed data and programs into it.

Let me simplify the basic requirements of information processing, speaking of verbal or non-numeric data:

(a) to distinguish what is worthy to be saved from what does not deserve to be recorded and retrieved;

(b) to identify in a document what is central, essential, and characteristic from what is just peripheral; which lead to

(c) to be able to retrieve everything concerning a given concept whenever it is needed; and consequently

(d) to summarize a document, and a set of documents, and a set of abstracts.

Automatic indexing, abstracting, and automatic translating may also be added.

To be able to do all that by computer, what we really lack is not more and better machines but much more data, much more knowledge to feed into the computer. Computer possibilities today are already much greater than our ability to use them, at least in language processing. Consequently, the new dimensions I referred to are concerned with our knowledge of the structure of our language and of our use of language. We lack much data and too many facts: we still know too little about our language to be able to make greatly significant increases in the use of computers. Verbal information, both written and oral, is a great mass of small elementary units, and this is precisely for what the computer is built. But the computer performs a single operation on any individual bit, one at a time. On the other hand, the mind of man processes information both as individual units and as a whole, subsuming and organizing the many as many into the one as one – many parts in one whole. As a consequence, when a man reads a sentence or a paragraph or a chapter, he understands much more than the individual words one by one: he understands the whole as a unit. And when this man wants the computer to perform operations implying the whole as a whole, he first has to find out the individual units (i.e. graphemes or grapheme distributions) in the text written on tape which express as symbols the whole as such, *if there are any*. If there is none, man has to add proper symbols to the text when transcribing it onto tape, expressing it precisely.

In my opinion, the new dimensions of the electronic era can be summarized to find out what physical entities or situations in a text express the semantic content of the whole sentence as such, if there are any.

In cybernetics (e.g. artificial intelligence), linguistics (e.g. the problems of presuppositions), and, to a greater extent, the science of scientific documentation, that point is being consciously attacked. I had to face that problem in my own work. In fact, one of the ways for controlling text redundancies was to identify the 'syntagmas', i.e. strings of words to be processed as one keyword. At some time in the future we expect that a computer will be able to select and summarize documents if we are able to attain that deeper insight into human information to which I referred.

Let me point out one consequence arising from the above. At the starting point of a new era there may be the temptation to ask the new techniques to do things in the same way as before. See, for example, some recent literature expressing critical remarks on computer use.[3] My statement is confirmed that using the computer to prepare concordances, for example, with the same format and the same features as before is a poor use of a computer. I feel sympathetic to anyone in scholarly research who still thinks of using a computer just to do things easier and faster. The processing of my *Index Thomisticus* took one million man-hours for much less than five thousand machine hours. In language processing the use of computers is not aimed towards less human effort, or for doing things faster and with less labour, but for more human work, more mental effort; we must strive to know, more systematically, deeper, and better, what is in our mouth at every moment, the mysterious world of our words.

In transcribing a text onto tape, some scholars may still think of writing just the graphemes printed in the text and nothing else; the result is certainly acceptable, as something is better than nothing. But taking into consideration the trends of research and development in computerizing information, I strongly recommend that anyone punching a text onto cards or tape should invest more effort in the project by adding additional codes to it, symbolizing at least some of the various typologies of the sentences and at least some basic semantic categories of the words. It is only by starting to take some step that we can hope to reach the challenging goal.[4]

Notes

1. This is a development of some of the statements made by the author at the 14th International Congress of Historical Sciences, San Francisco, 22-29 August 1975, when asked to comment on the report of Professor R. Morsey and Professor F. P. Kahlenberg.

2. See, for example, K. Gärtner and R. Wisbey, 'Zur Bedeutung des Computers für die Edition Altdeuscher Texte', in *Kritische Bewahrung, Beiträge zur deutschen Philologie*. Festschrift für Werner Schröder zum 60. Geburtstag (Berlin: Erich Schmidt, 1974), pp. 344-56.

3. W. Ingram, 'Concordances in the Seventies', *CHum*, 8 (1974), 273-7; J. Virbel, 'Compte-rendu d'une table-ronde du C.N.R.S. sur le thème: "Procedures d'analyse et méthodes de validation dans l'étude des donnêes textuelles"', *ALLC Bulletin*, 3 (1975), 172-3;
R. W. Ackerman, Review of '*A Concordance to the Works of Sir Thomas Malory*', edited by Tomomi Kato (Tokyo: University of Tokyo Press, 1974): in *CHum*, 9 (1975), 44-5;
S. Lusignan, Review of '*Thesaurus Bonaventurus*', edited by J. Hamesse (Louvain, 1972): in *Speculum*, 50 (1975), 311-14.

4. R. Busa, 'The Computer in Theological Research', in *Fachsprache – Umgangsprache*, edited by J. S. Petöfi et al. (Kronberg/Ts, 1975), pp. 403-24;
W. K. Koch, 'Some Critical Remarks on the Use of Computers in Linguistic and Literary Research', pp. 425-37;
R. Busa, 'Address: ALLC 1974 Annual General Meeting', *ALLC Bulletin*, 3 (1975), 56-6;
C. Montgomery, 'Linguistics and Information Science', *Journal of the American Society for Information Science*, 23 (1972), 195-219.

CONTENTS