1-1-1999

Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance

Thomas Nelson Winter
University of Nebraska-Lincoln, twinter1@unl.edu

ROBERTO BUSA, S.J., AND THE INVENTION OF THE MACHINE-GENERATED CONCORDANCE[1]

"The use of the latest data-processing tools developed primarily for science and commerce may prove a significant factor in facilitating future literary and scholarly studies."

— Paul Tasman

IBM Journal of Research and Development (July, 1957), p. 249.

Here in 1999, Classicists, both as teachers and as scholars, are equipped to appreciate that Tasman's 1957 prophecy was, if anything, an understatement.

Classicists, both as teachers and as scholars, are enjoying the fruits of the ongoing technological revolution. Classics teachers can now, with a few keystrokes, produce a customized concordance to any literary work we teach; we can generate word-lists for our students in frequency-of-occurrence order for efficiency-guided vocabulary study. The Internet has become a world-wide burgeoning encyclopedia at our keyboard: we can pull in texts off the Internet—or scan them ourselves—and run search packages on them; we can pull in photos of vases, or of archaeological sites and with a few more keystrokes turn them directly into transparencies for classroom use.

Classicists, as well as any, know that Tasman got it right.

Given the advances in these "data-processing tools developed primarily for science and commerce," as well as the work of Ted Brunner and Lucy Berkowitz (the *TLG*, the *TLG Canon*) and of David Packard (the programmed concordance, Ibycus, the Packard Humanities Institute) and Stephen Van Fleet Waite (LOGOI Systems, now in PHI), research work which even in 1980 required a mainframe computer can now be done at a desktop. In sum, the history of computers and the humanities has been all along a matter of broad-ranging scholars seeing opportunities in the newest commercially available data-handling equipment.[2]

It wasn't always this way. This is a story from early in the technological revolution, when the application was out searching for the hardware, from a time before the Internet, a time before the PC, before the chip, before the mainframe. From a time even before programming itself.

[1]We find in 1990 "About 30 years have passed since computers were first introduced into the humanities," Giacinta Spinosa, "Introduction," *Computers and the Humanities*, 1990, p. 349. The 1960 terminus is essentially right, which makes it the more important to look at—and appreciate—the work in the decade before it.

[2]A good example is Packard's development of the Ibycus from a Hewlett-Packard computer.

Let us step back by reviewing a series of key dates in computing, and humanities computing. Some entries below are for products, some are for tools.

1989 World Wide Web. Tim Bemers-Lee/CERN
1985 Perseus Project planning starts.
1980 Ibycus (mini-mainframe). David Packard.
1974 Through 1980—*Index Thomisticus:* (hard-copy, 56 voll.) R. Busa/CAEL
1972 TLG Planning Committee.
1968 David Packard's program-produced *Concordance to Livy.*
1967 Thomas Aquinas text card-punching completed. R. Busa/IBM
1957 (Published) Dead Sea Scrolls machine-readable. R. Busa/IBM
1957 Magnetic-tape assisted Bible Concordance. J. Ellison/Remington Rand.
1957 FORTRAN made public.
1953 Computers made public (IBM ships IBM 701)
1952 Programming invented.
1951 Machine-generated concordance. R. Busa/IBM
1946 ENIAC (electronic tube computer). U.S. Army.
1943 MARK I, IBM/Harvard, electronic relay computer.
1890 U.S. Census recorded on punch-cards; Hollerith founds IBM predecessor.
1884 Punch-cards. Herman Hollerith patent.

To start with the 1951 entry is essentially to start with the beginning of humanities computing. For further setting and contrast, the U.S. Bureau of the Census installed the first UNIVAC in 1951, and as noted above, IBM delivered its first computer, the 701, in 1953.[3] With regard to the above setting, how could you do a machine-generated concordance in 1951?

To begin to answer, Tasman's 1957 prophecy was no shot in the dark. His view of the future was a projection from his recent past. Thomas J. Watson, Sr. had assigned him in 1949 to be IBM liaison and support person for a young Jesuit's daring project to produce an index to the complete writings of St. Thomas Aquinas [Ref. 176, p. 84].

First, Tasman's thesis, as subsequent history turned out, was a huge understatement, and second, it essentially defines the first large invention of Father Roberto Busa, S.J., namely, to look at "tools developed primarily for science and commerce" and to see other uses for them. As will be seen, this was a case of fortune favoring the prepared mind. Redirecting

[3]Fisher, Franklin, et al. *IBM and the US Data Processing Industry: an Economic History,* New York, 1983, p. 8 and 12.

ROBERTO BUSA, S.J.

business record-keeping equipment[4] to the service of philological scholarship, he essentially invented the machine-generated concordance, the first of which he had published in 1951.

Father Busa, of course, is best known as the producer of the landmark 56-volume *Index Thomisticus*.[5] As he began this work in 1946, and produced a sample proof-of-concept, machine-generated concordance in

[4]The Computer-Tabulating-Recording Company (C-T-R), the conglomerate that became IBM, initially marketed meat-cutters, cheese-cutters, and grocery-store scales, in addition to registers and tabulators. See Rogers, William, *THINK, a Biography of the Watsons and IBM*, New York, 1969, p. 80.

1951, his professional life spans the entire computing chapter in the history of scholarship. Emphasis in this article will be on the early steps.

He announced the need for such a project in his dissertation, and made a presentation of the need for an *Index Thomisticus* at the International Congress of Philosophy, Barcelona, October, 1948 [Ref. 3, p. 425]. The announcement of the project in *Speculum* [Ref. 3, p. 425] starts us at the handmade level of the *Thesaurus Linguae Latinae*, and envisioned a work in two parts:

> a) A general file of all the words appearing in the works of St. Thomas. Each card in the file will contain in the upper left hand corner, a specific word used in the text of St. Thomas. Below that will be given the exact reference to the place in his text where the word appears, with a quotation of the sentence in which it is found. Such a file would need about thirteen million cards...
>
> b) Indices and Concordances to be drawn from such a file.

No hint of the revolution to come, except an appeal to scholars for

> any information they can supply about such mechanical devices as would serve to achieve the greatest possible accuracy, with a maximum economy of human labor. (Father Busa has been in contact with IBM in New York, the RCA laboratories in Princeton, the Library of Congress, and the Library of the Department of Agriculture, in Washington.)

Like all good projects, this one began with a question: What is the metaphysics of *presence* in St. Thomas Aquinas? Combing for *praesens* and *praesentia*, he realized that such words were peripheral, and, however unfortunately, Saint Thomas's doctrine of presence is linked with the preposition *in*!

Inquiring what St. Thomas meant by "presence," the young Roberto Busa realized that we must also study the way function-words affect meaning-words. To study the significant phrase "in the presence" he needed the shades of "in." His dissertation, defended in 1946, was essentially founded on a handmade Thomistic Concordance, essentially complete, but with one entry.

He had made 10,000 hand-written cards.

This project had results quite beyond the theological and philosophical value of his findings published in the first entry of the select bibliography below. Deeming it necessary to learn what significance

[5]Roberto Busa, S.J., *Index Thomisticus: Sancti Thomas Operum Omnium Indices et Concordantiae*, Stuttgart, 56 voll, 1974–1980. This is now available, of course, on CD-ROM. References to other works of Fr. Busa are keyed to the select bibliography that accompanies this article.

words have in an author's mind before attempting to gain insight into an author's conceptual system, he envisioned a concordance of all the words of St. Thomas Aquinas *including* the conjunctions, prepositions, and pronouns, a vision which required dealing with 10,000,000 words: all phrases broken out, each phrase copied over once for every word within the phrase, the lemma indicated on each of these cards, then sorted.... Having dealt with "grand games of solitaire" with 10,000 cards, he knew the scope. In short, he saw a need for the impossible.

Knowing this was impossible, he began a search for mechanical assistance. In the United States, mechanical manipulation of cards had already become a major industry. This industry was based on Herman Hollerith's patents, and the Hollerith patents were owned and developed by IBM.

> Hollerith, a graduate engineer from Columbia University, had developed a system of recording statistics for the U.S. Census Bureau back in 1890 by punching holes in sheets of paper. By adapting techniques he observed in the player piano and the Jacquard weaving loom, which was regulated by cards in which holes were punched to represent a pattern to be woven on the loom, Hollerith had, in fact devised the forerunner to the ubiquitous punch card with which IBM changed the world.[6]

Father Busa's survey of the mechanical record-keeping resources took him to the United States in 1949. His coast-to-coast search took him to about twenty-five American universities, a tour which led him to IBM and to an audience with Thomas J. Watson.

Watson, known to business history as the founding genius of IBM, was also a leader of the board of trustees of Columbia University, and had recently been quite busy getting Columbia to recruit Dwight D. Eisenhower as University President.[7] Given the requirements of the project, and the card-machine resources under IBM patent, the meeting was crucial, a make-or-break moment.[8]

[6]Rogers, William, *THINK, a biography of the Watsons and IBM*, New York, 1969, p. 69.

[7]Rogers, pp. 203–208.

[8]The situation of the then-contemporary data-processing industry has implications. That *Varia Specimina* is the world's first machine-generated concordance approaches the certainty of a theorem in geometry.

(a) Suppose the contrary, i.e., that there was a machine-generated concordance before 1951.

(b) If it was done before 1951, it was done on punch-card manipulating machines.

(c) If it was done on punch-card manipulating machines of that era, it was produced with the full cooperation of IBM.

(d) But this *is* the project done with that cooperation.

Item (b) above is a certainty. Item (c) above, is not fully as certain: Remington Rand equipment is an outside possibility. The US Bureau of the Census lodged an anti-trust suit against IBM before the war for monopolizing the punch-card

Father Busa recalled the meeting as follows:

> I knew, the day I was to meet Thomas J. Watson, Sr., that he had on his desk a report which said IBM machines could never do what I wanted. I had seen in the waiting room a small poster imprinted with the words:
>
> "The difficult we do right away; the impossible takes a little longer."
>
> (IBM always loved slogans.) I took it in with me into Mr. Watson's office. Sitting down in front of him and sensing the tremendous power of his mind, I was inspired to say: "It is not right to say 'no' before you have tried." I took out the poster and showed him his own slogan. He agreed that IBM would cooperate..."provided that you do not change IBM into International Busa Machines." [Ref. 176, p. 84]

The first product of this alliance was formed partly by the limitations of the then current IBM equipment. It would manage only eighty characters on a card. Eighty characters? Father Busa realized this would allow nothing longer than, say, a line of hendecasyllabic poetry. And a lemma.

So he did the poetry of St. Thomas Aquinas. His 1951 machine-generated and machine-printed concordance served as a proof-of-concept exercise. How was it done?

In the preface to the *Varia Specimina*, Father Busa summarizes the essential stages of work for generating concordances:

1. Transcription of text, broken down into phrases, on to separate cards;
2. multiplication of the cards (as many as there are words on each);
3. indicating on each of the resulting cards the respective entry (lemma);
4. selection and alphabetization of all cards purely by spelling; and
5. after the intelligent editing of the aphabetism, the typographical composition of the pages for publishing.

data-processing industry. In that suit, it was revealed that IBM had a no-competition agreement with Remington Rand, whose card-machines used mechanical pins instead of electric brushes to read the punch-outs. Remington in 1935 held the 15% of the market left over from IBM (Rogers, 129–130). The question becomes, was there a Remington Rand-associated concordance project? Not with punch-cards. The first Remington Rand-associated concordance project appears to be John W. Ellison's Bible Concordance of 1957. It was done with the new magnetic tapes.

Record-management in 1949

At the time Father Busa entered the scene, only the second of these concordance-tasks, the multiplication of cards, was being done mechanically. The *TLL* was using the services of a copying bureau; Professor Roy Deferrari, Catholic University of Washington, was using electrical typewriters which could make many copies; Professor P. O'Reilly of Notre Dame University, South Bend, Indiana, had each side of the text-page repeated as many times as there were words thereon. [Ref. 4, p. 22]

Something was needed which could do all five. Promising was the Rapid Selector, invented by Vannevar Bush and developed by Ralph Shaw. Father Busa saw it operating in 1949 at the Library of the Department of Agriculture in Washington, D. C. It simultaneously ran two microfilm reels, one with text, the other with coded symbols for the words of that text, and photographed the pages exhibiting the targeted keywords. It examined 10,000 microfilmed pages per minute, instantaneously rephotographing the found pages on another microfilm! [Ref. 4, p. 22] Though impressive, this looked like a blind end—all the coding had to be done by hand! And it could not be adapted to automated printing.

One cannot but note that Father Busa knew the nature of the task and knew what he was looking for. In the array of business machines in the post-war IBM inventory, the punchcard electro-mechanical accounting machines looked more promising than anything he had seen at the American libraries or in other indexing projects.

What were these machines? How did they work? What was Father Busa looking at—and later dealing with? No one has set this forth better than Dr. Cuthbert Hurd, testifying in US vs IBM, the antitrust suit brought by the Justice department in 1969.

> The components of punched card equipment included brushes which would detect...a hole in a punched card and which then produced an electric signal, commutators which divided an electric signal into a number of timing intervals, relays which opened and closed...mechanical devices for punching holes in cards, and mechanical printers. Relays could be opened and closed a few dozen times a second and were subject to unreliable operation because they were mechanical and because of dust particles.... IBM had built a variety of machines using these components, including a *key punch, verifiers, interpreter, reproducer, gangpunch, collator, tabulator, sorter, and calculator.*[9]

This is the array in which Father Busa managed to see the *Index Thomisticus.* Computer? Not yet. Obviously applicable to Humanities research? Not exactly. But to continue Hurd's testimony:

[9] In Fisher, p. 13.

> These devices were controlled by control panels, or "plug boards...." Such a control panel might measure three feet by two feet and contain perhaps a thousand holes. Each machine had a different control panel...manual intervention was the key and because of manual intervention and because of the mechanical nature of the devices, the results were slow and unreliable. Consequently, there was a sharp limit on the size and kind of application or tasks that could be performed.

Father Busa wrote in 1951, in the preface to his sample machine-generated concordances:

> What had at first appeared merely as intuition can today be presented as accomplished fact: The punched card machines carry out all the material part of the work in points 2, 3, 4, and 5. [Ref. 4, p. 22]

Varia Specimina

Varia Specimina is a remarkable document for two essential reasons. First, it was the demonstration that the punch card accounting machines could do all five tasks essential to the production of a concordance. Second, in describing the process in the preface, Father Busa gave us a virtual time machine to see this process in its first generation.

First trials were done on one of Dante's *Cantos*. Each of the 136 canto lines was punched (as position-coded holes) onto a card at the first machine, the automatic punch. This was operated by a keyboard like a typewriter, and was the only human input. Father Busa points out that an error at this level would thus be perpetuated, but from this point on, error-free would remain error-free:[10] no new "typos" would have a chance to creep in! [Ref. 4, p. 24]

Proofreading was done by machine also, the Collator. Each line was typed in twice. If the collator sees no difference, both are correct; if there is a difference, a typographical error is flagged, and the line was re-punched. The rest was largely a matter of overseeing the machines. At the next station, the Record Interpreter machine printed what was 'written' in the holes. At the next station, the Reproducer made a copy of each card, plus, at the side, the first word of the line. Then a second copy, with at the side, the second word of the line. Finally, there were as many cards as words in

[10]This was an abiding theme for the entire project. In their write-ups of the time, both Busa in Europe and Tasman in New York used as a facing card illustration Thomas's words—reminiscent of Aristotle—QUIA PARVUS ERROR IN PRINCIPIO MAGNUS EST IN FINE.... It made a good sample punch card.

the text. In other words, each line was multiplied as many times as the words it contained. At this point, alphabetizing was a matter of feeding the Sorter Machine [Ref. 4, p. 26], and if a researcher were to drop the cards, getting them in correct array would be (ideally) a matter of simply putting them in the Sorter again.

Most of the manual intervention occurred here, as the business machine did not begin the second word in the same column on all cards, or the third, or the fourth.... The language itself was trouble-causing. E.g. in the Italian *o, ebbi, avrei*, are all forms of the same word; *andiamocene* is three words presented as one; in Latin *mortuus est* is (functionally) one word presented as two. When the necessary manual intervention was done, the last stop was the Alphanumeric Accounting Machine, or Tabulator. This retranscribed the words in the holes in the cards into the letters and numbers. A print-out, in sum. As Father Busa put it in his preface: "The concordance which I am presenting as an example is precisely an off-set reproduction of tabulated sheets turned out by the accounting machine." [Ref. 4, 28]

The prospectus from 1951 marked a turning point, and is worth preserving:

> "It is at all events certain that from now on the history of concordances will no longer have to record figures like those of the past: 500 Dominicans—can it really be true?—employed by Hugh de St. Cher in 1200 in Paris for the first Biblical Latin Concordance; fifty monks occupied in preparing the Biblical concordances organized by the Benedictines in 1700; five German Universities cooperating to set in order by hand the ten million cards of the *TLL* at the end of the last century.
>
> "Today it will suffice that man's hands transcribe by typing the entire text on the Punch, control its accuracy either perusing the cards which the Interpreter will have made legible or using, as I have mentioned, the Verifier or the Collator; this done, the Reproducer, the Sorter and the Accounting Machine and Collator will take entire care of the remaining material part of the work; in a few days the philologist, with the mere help of a technical expert for the care of the machines, will have in hand the general card file and the final proofs corrected for the printer, certain of an accuracy which could never have been guaranteed by the cooperation of man's sensorial and psychical nerve centers." [Ref. 4, pp. 34, 35]

The 1951 demonstration that accounting machines could generate a concordance offered six tools for scholars working with the poems of St. Thomas:

1. Words and their frequency, forwards
2. Words and their frequency, written backwards
3. Words set out under their lemmata (eg, *aemulis* under *aemulus aemula aemulum*), with frequency
4. The Lemmata
5. Index of the words
6. Keyword in context Concordance

The project was done in IBM offices in Milan, where Father Busa started his own punching and verifying department.

In 1954, he started a training school for keypunch operators.

> "The success was excellent: industries wanted to hire them before they had finished the program. *Their training was in punching and verifying our texts.*" [Emphasis added.] The school was continued until 1967, when the text-punching was finished. He then moved his operations to Pisa, then in 1969 to Boulder, Colorado, and in 1971 to Venice, "wherever IBM provided computer time." [Ref. 176, p. 85]

In the 1951 project, no computers, no programming. Though there is no attempt here to recount the history of the computer, it should be noted for context that the first electronic (vacuum tube) computer was developed for the U.S. Army and operational in 1946, the ENIAC. Programming, the storing of operating instructions in the machine so it could operate independently, was invented in 1952, and the invention is usually credited to John Von Neumann of the Atomic Energy Commission.[11]

Back in the United States, Benjamin D. Wood, a pioneer in educational measurement who had to oversee the scoring of 35,000 examinations a year, also got an appointment with Thomas J. Watson senior. The immediate result was a convoy of trucks bearing card machines to Columbia University—which was never billed for them. In reviewing the work which these machines were doing, and interrupting a gush of grateful superlatives, Watson asked Wood "What's *wrong* with our machines?" He was not ready for the answer.

Wood replied they were too slow. "Ultimately these machines, which are now electro-mechanical, since they operate on electricity, are going to operate at the speed of light, 10,000 times the speed at which they function now." Watson's biographer reports

[11] *THINK*, p. 175.

"A gleam appeared in Watson's eyes...."[12]

Only the tiniest fraction of the machine time per card was spent reading its punches. Almost all was in the physical manipulation of the card. The now-familiar mark-sense machine scoring was not possible. Carbon from pencil marks had an electrical resistance, 500–5,000 ohms, varying by a factor of ten. A Michigan school teacher, Reynold B. Johnson, invented a scoring machine with such high resistance that the pencilmark resistance was not consequential. IBM bought it.[13]

The Transition of the Later Fifties

Programming in the *Varia Specimina* project had been a matter of planning the process. The process was one of marking instructions on the text or on a transparent overlay on the text for the keypunch operator, and then overseeing the machines and overseeing the transport of cards from machine to machine. When we next see a report of the hardware and software behind the project, we see mark-sensing, a transition to magnetic tape, and a mention of programming.

Learning in the press (apparently in 1954 [Ref. 25, p. 373n]) of John W. Ellison's Bible Concordance work with Remington Rand magnetic records, Father Busa then met him in person, shook his hand saying "You are a great ally of mine," then immediately went back to IBM, where he asked "See what Remington is doing?" [Ref. 176. p. 85]

Punch-card data-processing, though miraculous compared to hand-work, was slow: Here is the view from 1958: "Mechanical alphabetizing requires two passes of the cards through the machine for each column sorted. Thus sorting 100,000 cards containing words of ten letters means in effect, passing 2,000,000 cards through the machine. Depending on the model sorter used, from 30,000 to 60,000 cards per hour can be sorted. Therefore it could take from thirty-five to sixty-five hours, approximately, to accomplish the alphabetization. In other words, the machine would alphabetize from 1,500 to 3,000 words of ten letters in an hour." [Ref. 25, p. 361]

Even this, of course was appreciated. Paul Tasman produced this comparison in 1957, when the 2000 pages of the *Summa Theologica* were represented by 1,600,000 word-cards:

1. Manually, three persons, 20,000 hours.
2. Punch-card method: three persons, 1,000 hours.
3. "Large-scale data-processing method," one person working 60 hours.[14]

[12] *THINK*, p. 138.
[13] *THINK*, p. 139.

The definition of what he meant by "Large-scale data-processing method" glimmers in the next sentence: "This, of course, is exclusive of preparation and *programming* time." [emphasis added] Programming is now in civilian use, but it is still in card—or card-analog—management.[15]

The cards are now, figuratively speaking, two-edged. Machine punched for the one, mark-sense capable for the other. The same machine can now read both holes and pencil-lead. That is, the cards enable built-in, machine readable human intervention—for which, as we will see, there is always great need.

Tasman gives the background to Father Busa's Dead Sea Scrolls project:

> In this application, the major objective is compactness of magnetic tape files and the speed at which tapes can be read, written, and printed.
>
> Each word card...is initally converted to magnetic tape in blocks of records 80 characters long. The process is broken down into four runs of the IBM 705. Run 1 deals with sorting and inverting each word in memory. Run 2 deals with creating the frequency count of the stored words and their summarization. Run 3 deals with a merging and grouping of the different word tapes with the entry word tape. Run 4 provides for collating each word tape with the original phrase tape.[16]

Much of the actual "programming" was still in the design of the card, and the essential over-all program was still visibly the process for producing a concordance laid out by Father Busa in *Varia Specimina*. I venture to summarize the 1957 algorithm, from Tasman's section "Specific phases of automation in the literary analysis of the *Summa Theologica* of St. Thomas Aquinas":

1. The scholar analyzes the text, marking it with precise instructions for card punching, and phrase limits.
2. A clerk [a student in the keypunch operators school] copies the text. The phrase is now holes in a card, preceded by a reference, and a serial number. A second clerk retypes the phrase on the same card [apparently a change from 1951]. Cards with discrepant phrases are found by a checking machine, and replaced.

[14]Paul Tasman, July, 1957 *IBM Journal of Research and Development*, p. 256.

[15]The first FORTRAN system was released in 1957, for the IBM 704. This was the year of Tasman's report/prospectus. FORTRAN was, of course, for numerical computation. See S. Ramsden, University of Manchester, http://www.man.ac.uk/hpctec/courses/Fortran90/Fortran90_4.html

[16]Tasman, p. 256.

3. From the phrase-cards the machine [the IBM 705?] does two jobs: a) produces the word cards, and b) produces a complete copy of the text, phrase by phrase. In various zones of the card, there are encoded 1) the reference, 2) the first letter of the preceding word and the first letter of the following word, 3) the ordinal number denoting the word's position in the text, 4) a characterizing mark.[17]
4. Form cards: the machine counts and eliminates all word duplicates. In this stack, there is left only one card for every graphically different word, and a record of its total occurrences. At this stage, *sum* and *est*, for example, are still different entries.
5. The scholar converts the form-cards into the entry cards.

At this point in the algorithm, the scholar's task is two-fold, and Herculean. He must first examine the form cards and group all the different forms of the same semantic unit under the one word which will serve as the entry, or lemma. He must then split all homographs into their respective lemmas. For example, in English, *lead* the metal must be separated from *lead* the verb; in Latin, *amor* the noun from *amor* the passive first person singular verb form, or *labor* the noun "work" from *labor* the deponent verb "I am slipping."

This is morphology on the practical level. It is intense work. Professor Busa has written a book about it which will reward anyone interested in Latin. It is *Inquisitiones Lexicologicae.* [Ref. 258]

6. Where not already done, all groups of cards are "interpreted." This means the machine prints on top of each card whatever information is on or in the card.
7. "Information on any desired groups of these cards can now be printed on sheets, in brochures or books or on other cards. Valuable tools in philological research like an index verborum or concordance are available without further scholarly effort."

The work had interesting consequences for text-restoration. In Dead Sea Scroll experiments, "up to five consecutive words have been 're-written' by the data-processing machine in experimental tests where the words were intentionally left out of the text and blank spots indicated."[18]

[17]The characterizing marks? Some examples:
/ marked a passage where St. Thomas refers to other passages of his own work;
marked the words of another author quoted by St. Thomas;
□ marked phrases not entirely Thomistic as St. Thomas recorded, discussed, or refuted the doctrine of others.

[18]Tasman, p. 256n.

Conclusion

More generally, Tasman foresaw a new era coming from the techniques developed in the Busa/IBM project:

> The indexing and coding techniques developed by this method offer a comparatively fast method of literature searching, and it appears that *the machine searching application may initiate a new era of language engineering.* [emphasis added] It should certainly lead to improved and more sophisticated techniques for use in libraries, chemical documents, and abstract preparation, as well as in literary analysis.[19]

In conclusion, Father Busa, with IBM's enabling help, was at the pivot point (or was the pivot point) between handmade scholarly tools and machine-made scholarly tools. His impossible dream of 1949 is the reality of 1999. We are *in* Tasman's "new era of language engineering." As for the *Index Thomisticus*, the goal which motivated the transition, it has been completed and has undergone transformations from punch-cards to tapes to books to CDs. There is a growing bibliography of theological and linguistic work founded upon it.[20]

Appendix 1: A Challenge for Further Work

But, in the words of Father Busa himself in 1990, "Our generation has not done everything: for the young people there are still immense open spaces" [Ref. 243, p. 339]. He posed eight challenges/desiderata:

1. The automatic linking of syntagmata—sets of more than one word that should be treated as a unit, e.g. "surnames and names, the compound form of verbs, expressions such as 'clothes horse.'"

2. The automatic identification of the word to which a pronoun refers.

3. The detection of tacit words.[21]

[19]Tasman, p. 256.

[20]Some may be sampled in Paolo Guietti, "Hermeneutic of Aquinas's Texts: Notes on the *Index Thomisticus*," *The Thomist*, 57, (1993) pp. 667–686, where one may also see a knowledgeable view of the *Index*.

[21]I take "tacit words" to mean words which are sufficiently implied, but not present as, for instance, in the Greek geometric texts, the word for "line" can be represented by simply the feminine article or feminine adjective; "point" by the neuter.

4. The creation of a general conspectus—within the confines of a precise universe—of all words which are potentially homographic.

5. For every homographic word, or at least for each type, the recognition of what in the context is specific to one or the other of its values.

6. The recognition of all elementary and direct grammatical, i.e., semantic connections. For instance, which words in a given sentence are used as adjectives, and to which nouns do they refer? what is the object of a given verb?

7. The definition of the logical function of every word or clause; whether a subject, an object, a verb, or a complement.

8. Finally, the formalization of a logical train of thought winding through many paragraphs; this would permit automatic indexing and automatic abstracting.[22]

Appendix 2: Select Bibliography

The publications of Father Busa outside of the *Index Thomisticus* form a corpus of more than 200 entries, of which the following is a significant selection.

1. *La Terminologia Tomistica dell' Interiorità. Saggi di metodo per una interpretazione della metafisica della presenza,* Milano, Bocca, 1949, pp. 280.

3. "Complete Index Verborum of St. Thomas Aquinas," in *Speculum—A Journal of Mediaeval Studies*, XXV, (Jan. 1950), pp. 424–425.

4. *Sancti Thomae Aquinatis Hymnorum Ritualium Varia Specimina Concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate,* Milano, Bocca, 1951, pp. 180.

[22]I have shortened or paraphrased some of these desiderata. They are in Ref. 243, p. 341. Father Busa credits Peter Luhn of IBM with formulating challenge number 8 in the early 1950s, and also credits him as the introducer of the KWIC Index. Finally, a personal note: working towards desideratum number five has, coincidentally, been the core of my career as a researching Latinist since 1980—which is the vintage of the IBM 360 mainframe in the background of Father Busa's picture.

7. "Mechanisierung der philologischen Analyse," in *Nachrichten fur Dokumentation*, vol. 3, n. 1 (Mar. 1952), pp. 14–19.

12. "La Cibernetica," in *Ingegneria Meccanica*, Milano, III-2 (Feb. 1954), pp. 5–10.

25. "The Use of Punched Cards in Linguistic Analysis," in *Punched Cards—Their Application to Science and Industry*, 2nd edition, R.S. Casey et al. ed., New York, Reinhold 1958, pp. 357–373.

28. "All non-biblical Dead Sea Scrolls published up to December 1957 have been indexed," in *Sacra Pagina, Miscellanea Biblica Congressus Internationalis de Re Biblica* (1958) Paris, Lecoffre 1959, pp. 7–12.

30. "Erlauterungen zu den lexicographischen Arbeiten zu Goethe," Farbenlehre, Bd. 3 Kolloquium: Maschinelle Methoden der literarischen Analyse und der Lexicographie. Tuebingen, 24/26 Nov. 1960, pp. 36 (litografia).

35. *Atti del Convegno: Linguistica e Industria Oggi* a cura di R. Busa S.J. Milano, 19 Genn. 1962, pp. 75 (ciclostilato).

40. L' Analisi linguistica nell' evoluzione mondiale dei mezzi di informazione in *Almanacco Letterario Bompiani 1962*, Milano, 1962, pp. 103–107.

41. Stellungnahme in "Zur mechanischen Sprachuebersetzung. J. J. Becher, Allgemeine Verschlusselung der Sprachen" von Prof. W. G. V. Waffenschmidt, Kohlammer Verlag Stuttgart, 1962, pp. 45–52.

45. "An inventory of fifteen million words," in *Literary Data Processing Conference Proceedings, 9/10/11 Sept. 1964*, Modern Language Assoc. New York, 1965, pp. 64–78.

47. "World Evolution in Information Processing and its Influence on Linguistic Research" (in Hebrew), in *Third World Congress of Jewish Studies, Jerusalem 1961*, Jerusalem, 1965, pp. 47–49.

48. Die Elektronentechnik in der Mechanisierung der sprachwissenschaftlichen Analyse (Traduzione in russo dell' articolo del 1957) in *Avtomatizatsiva v lingvistike*, L. N. Zasorina, Nauka, Moscow, 1966.

53. "Saggio esplorativo di automatic abstracting," in *Convegno Nazionale T. D. 66: La riduzione Concettuale dei Documenti. Torino 14.12.1966.* Atti a cura del CSAO Torino, 1967, pp. 39–47.

63. "Erreurs Humaines dans la Preparation de l'Input pour Ordinateurs," in *Les Machines dans la Linguistique. Colloque International sur la mecanisation et l'automation des recherches linguistiques*, Accademia, edition de l'Academie Tchecoslovaque des Sciences, Prague, 1968, pp. 279–284.

66. *Actes du Seminaire International sur le Dictionnaire Latin de Machine Pisa 27/28/29 Mar. 1968*, redigé par Roberto Busa, S.J., pp. 176. (Appeared as supplement n. 2 to vol. V of *Calcolo*, Inac, P. le delle Scienze 7, Roma) 1968.

69. "L' Instrumentation electronique dans les recherches linguistiques," in *Rencontre International de Mechanographie et Informatique, Lisbonne Oct. 1967*, Caixa Nacional de Pensões, 1970, pp. 737–747.

71. "Concordances," in *Encyclopedia of Library and Information Science*, Marcel Dekker, New York, vol. V 1971, pp. 592–604.

73. "The Impact of Cybernetics on the Humanities," in *Proceedings of the Jurema 1972: International Symposium on Cybernetics in Modern Science and Society*, ed. W. Muljevic Zagreb, 1972, pp. 13–24.

104. "Inner and Outer Information: Causality types in man-to-man information," in *Proceedings of the Jurema 1975 International Symposium on Cybernetics in Modern Science and Society*, ed. W. Muljevic, Zagreb, 1975, pp. 17–20.

107/108. *Concordantiae Senecanae, R. Busa S.J., A. Zampolli*, Georg Olms, Hildesheim, 1975. 2 voll., pp. IX+VII+1473+59+59+5R.
[*Father Busa's masterful Seneca concordance is probably the work by which he is known to classicists. Perhaps the best measure of its worth is that no one has ever tried to supersede it; one thinks of the competing, and flawed, concordances of e.g. Vergil and Manilius. In sum, its trustworthiness and ease of use have ensured that it remains a welcome* vade mecum *even in the age of CD-ROMs.* (Ed.)]

136. "Computer Processing of Over Ten Million Words: Retrospective Criticism," in *The Computer in Literary and Linguistic Studies: Proceedings of the Third International Symposium*, Alan Jones, R. F. Churchouse, edd. The University of Wales Press, Cardiff, 1976, pp. 114–117.

147. "Man-machine relationship in computerized linguistics," in *XXVII Convegno Internazionale delle Comunicazioni, Genova 9/12 Ott. 1979*, Istituto Internazionale delle Comunicazioni Genova, 1979, pp. 279–286.

149. "ORDO dans les oeuvres de St. Thomas d'Aquin," in *ORDO II Coll. Internazionale Lessico Intellettuale Europeo, Roma 7/9 Genn. 1977*, Ed. Ateneo-Bizzarri, Roma, 1979, pp. 59–184.

176. "The Annals of Humanities Computing: the *Index Thomisticus*," *Computers and the Humanities*, New York, 1980, pp. 83–90.

178. "Per S. Tommaso 'ratio seminalis' significa 'codice genetico': problemi e metodi di lessicologia e lessicografia tomisticha, in *Atti dell' VIII Congresso Tomistico Internazionale*, vol. 1, Città del Vaticano, 1981, pp. 437–451.

193. *Global Linguistic Statistical Methods to locate style identities: Proceedings*, ed. by R. Busa, S.J. Lessico Intellettuale Europeo, Ateneo, Roma, 1982, pp. 111.

195. "Trente ans d' analyse informatique de textes: où en est-on? et après?" in *Actes du Congres International Informatique et Sciences Humaines, Liège 18/21 Nov. 1981*, Ed L.A.SoL.A. Liège, 1983, pp. 135-148.

200. "De terminationum Latinarum statisticis mensuris ex Indice Thomistico," *in Hommage à Pierre Guiraud*, Les Belles Lettres, Nice, 1985, pp. 340.

206. "Informatica e Nuova Filologia," in *Lessicolografia, Filologia e Critica—Convegno Internazionale di Studi Catania-Siracusa 26/28 Apr. 1985 Atti*, Ed. Olschki Firenze, 1986, pp. 200.

208. "De Linguae Latinae flexivis terminationibus," *Revue Informatique et Statistique dans les Sciences Humaines*, XXI 1-4, Liège, 1985, pp. 53–66.

215. "L' originalité linguistique de St. Thomas d'Aquin," ed. A.L.M.A. XLIV-XLV, Brill-Leiden, 1985, pp. 66–90.

219. *Fondamenti di Informatica Linguistica*, Ed. Vita e Pensiero, Milano, 1987, pp. 412.

220. "Das Problem der Thomistichen Hermeneutik nach der Veroffentlichung des Index Thomisticus," *Miscellanea Medievalia*, vol. 19, Koln, 1988, pp. 359–364

222. "De phantasia et imaginatione iuxta S. Thomam," in *Phantasia-Imaginatio: V Coll. Internazionale Lessico Intellettuale Europeo-Roma 9/11 Gen. 1986 Atti*. Ateneo Roma, 1988, pp. 135–152.

225. "Inteligencia natural e inteligencia artificial," *Broteria*, Lisbon, 1988, pp. 260–268.

229. *Totius Latinitatis Lemmata - quae ex Aeg. Forcellinii Patavina Edizione 1940 a fronte, a tergo atque morphologice, opera IBM automati ordinaverat Robertus Busa S.J.*, Istituto Lombardo di Scienze e Lettere Milano, 1988, pp. XVI + 532.

230. "Procedures et resultats de la segmentation thèmatique des lemmes latins de l'Index Thomisticus," *Revue Informatique et Statistique dans les Sciences Humaines*, 24, 1–4 Liège, 1988.

243. "Informatics and New Philology," *Computers and the Humanities*, New York, 1990, pp. 339–343.

244. "Idea negli scritti di Tommaso D' Aquino," in *Idea - VI Coll. Internazionale Lessico Intellettuale Europeo - Roma 5/7 Genn. 1989 Atti*, Ateneo Roma, 1991, pp. 63–87.

248. "De Expressione apud S. Thomam," in *Littera Sensus Sententia - Studi in onore di C. L. Van Steenkiste, OP*, Ed. Massimo Milano, 1991, pp. 135–154.

250. "Cinguant' anni a 'bitizzar' parole," in *Convegno Internazionale sulla Storia e Preistoria del Calcolo Automatico e dell' Informatica - Siena 10/12 Sett. 1991 Atti Precongressuali*, Siena, 1991, pp. 72–82.

251. "Half a Century of literary Computing: Towards a 'new' Philology," in *International Conference on Current Issues in Computational Linguistics—Penang 11/14 June 1991 Atti Precongressuali*, Penang, 1991, pp. 84–95.

257. *Thomae Aquinatis Opera Omnia (cum hypertextibus) in CD ROM* Milano Editel 1991, 1 CD ROM + 64 pp.

258. *Inquisitiones Lexicologicae in Indicem Thomisticum*, Milan, with accompanying English tr. by Philip Barras, Milan, 1994.

THOMAS NELSON WINTER
University of Nebraska