Roberto Busa (1913–2011): Lemmatizing Aquinas Automatically

Bod, R.

DOI
10.1086/704849
Publication date
2019
Document Version
Final published version
Published in
History of Humanities

Roberto Busa (1913–2011): Lemmatizing Aquinas Automatically

Rens Bod, *University of Amsterdam*

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?[1]

Father Busa's short book with an excessively long title, *Sancti Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum: A First Example of Word Index Automatically Compiled and Printed by IBM Punched Card Machines*, is generally acknowledged as the starting point for processing texts with machines. Busa has been called the "founding father" of both humanities computing—today known as digital humanities—and computational linguistics.[2] Despite these triumphalist qualifications, Busa's book is rarely read, let alone anthologized. There may be several reasons: (1) the book is of a rather technical nature which is or was alien to the prevalent tradition in the humanities; (2) in the field of humanities computing, projects usually receive more attention than the writings about these projects; and (3) the English version of Busa's bilingual book is too literal a translation of the Italian and not a particularly pleasant read. Nevertheless, Busa's piece is a classic in the short history of the digital humanities in that it not only analyzes but also criticizes the application of computational tools in humanistic scholarship.

Roberto Busa (1913–2011), a native of Vicenza, entered the Jesuit order at age twenty. While working as a chaplain during the Second World War in the Italian army and later in the partisan forces, he simultaneously carried out research on the metaphysics of

1. Employed by Hugh de St. Cher in 1200 in Paris for the first Latin Bible Concordance. See Roberto Busa, *Sancti Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum: A First Example of Word Index Automatically Compiled and Printed by IBM Punched Card Machines* (Milan: Fratelli Bocca Editori, 1951), 32. The book contains forty-nine pages of text, of which the even pages are in English, and the odd pages in Italian.

2. Susan Schreibman, Ray Siemens, and John Unsworth, eds., *A Companion to Digital Humanities* (Oxford: Wiley-Blackwell, 2005); Matthew K. Gold, *Debates in the Digital Humanities* (Minneapolis: University of Minnesota Press, 2012).

History of Humanities, Volume 4, Number 2. http://dx.doi.org/10.1086/704849

"presence" in Thomistic philosophy. During that research, he wrote out by hand around 10,000 cards with the word "in" or a word connected with "in." In 1946 he graduated with a degree in philosophy at the Pontifical Gregorian University in Rome. In the same year, Busa planned the *Index Thomisticus*, which was aimed to be a tool for text searches within the 11 million word corpus of Thomas Aquinas's works. Busa started to look for machines for the automation of such a task, which he finally found at IBM, resulting in a collaboration that lasted over thirty years. The entire texts by Aquinas were transferred to punched cards, and a concordance program was written for the project. The intention was to produce printed volumes, of which the first was published in 1974. When the project was finished in the 1980s, it delivered fifty-six printed volumes. A web-based version was made available in 2005.[3]

In his book, Busa gives a description of the techniques and processes as applied to the compilation of concordances and *indices verborum* using IBM machines. Busa introduces technical phrases that are currently in use but that were new at the time, such as "running in *tab*." He compares the enormous labor devoted in the past to such works with the relative ease of the system under consideration. But Busa also criticizes the initial results of his project, such as the loss of accents, punctuation, and the distinction between capitals and small letters. These shortcomings were shared by all machines until 1950, as well as by the early mainframe computers in the 1950s, due to the high costs of computing memory. Thanks to Busa's project, IBM started to make efforts to deal with the full richness of texts. This was greatly helped by the move from mechanical to electronic machines, setting the stage for later applications. Computers had come to be no longer viewed as mere calculating machines but also as text-processing systems. Tools were developed to make indexes, lemmatizations, word counts, and concordances. It is the critical discussion of these tools that makes Busa's book valuable even today. Moreover, it gives us an insight into the practice of humanities computing in its first generation.

For example, Busa discusses a selection of poems by Thomas Aquinas and goes into the advantages and shortcomings of the IBM tools. He discusses the automatic construction of six indexes: an alphabetical index with the frequency of the words, a reverse index, an index of the various forms of each lemma (e.g., *aemulis* under *aemulus aemula aemulum*), a simple listing of the lemmata, a verbal index that gives every single occurrence of each word, and a concordance index that gives the entire line of the text in which the lemma occurs. Busa stresses the greatest advantage of his approach: the six indexes were compiled in a few hours by the IBM machines, a task that would take years if done manually. But he also notes the severe disadvantages, for example, when the tool does not recognize *que* as a separate lemma when it is not separated by a space from another

3. See http://www.corpusthomisticum.org/.

lemma (such as in *sanguinisque*, *trinoque*, or *unaque*). He also notes several other errors—for example, *genito* does not derive from *genero*, *generare*. These are problems every digital humanities student still runs into and struggles with when learning how to program tools for text processing. Tools must thus not only be used but also criticized and subsequently improved. Even though Busa and his team attempted to write computer software that could handle the problems above, none of the programs performed without errors, and the lemmatization of all 11 million words could only be completed in a semi-automatic way, with human beings dealing with word forms that the program could not handle. These shortcomings have yet to be overcome.

What then is Busa's legacy? It took a few decades before Busa's tools were used at a (relatively) wide scale in the humanities. This was because IBM's text-processing tools could run only on very expensive machines, such as the IBM 701 mainframe. The community of computational linguists and philologists started to grow only in the 1960s and remained very modest until the 1980s. With the introduction of the personal computer, textual tools, as well as commercial text processing systems, became more common. Nowadays, word-processing systems are part of the toolkit of every humanities scholar. Yet it was Busa who was able to breach the barrier between humanities and computer science—though he was not alone in this, as the contribution by Mario Wimmer in this issue has made clear. Busa also showed that the interconnection between these two fields—humanities and computer science—led to new problems that could not be foreseen when indexes were still created manually. Busa is therefore not only a founder of humanities computing but also of tool criticism.

FIRST PUBLICATION

Busa, Roberto. 1951. *Sancti Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum: A First Example of Word Index Automatically Compiled and Printed by IBM Punched Card Machines*. Milan: Fratelli Bocca Editori.

RECOMMENDED SELECTION FOR STUDENTS

Busa, Roberto. 1951. *Sancti Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum: A First Example of Word Index Automatically Compiled and Printed by IBM Punched Card Machines*, 1–25. Milan: Fratelli Bocca Editori.

RECOMMENDED READINGS

Bod, Rens, Jaap Maat, and Thijs Weststeijn, eds. 2014. *The Making of the Humanities*. Vol. 3, *The Modern Humanities*. Amsterdam: Amsterdam University Press.

Busa, Roberto. 1980. "The Annal of Humanities Computing: The Index Thomisticus." *Computers and the Humanities* 14:83–90.

Jones, Steven E. 2016. *Roberto Busa, S. J., and the Emergence of Humanities Computing: The Priest and the Punched Cards*. New York: Routledge.

Nyhan, Julianne, and Andrew Flinn. 2016. *Computation and the Humanities: Towards an Oral History of Digital Humanities*. Basel: Springer.

Schreibman, Susan, Ray Siemens, and John Unsworth, eds. 2005. *A Companion to Digital Humanities*. Oxford: Wiley-Blackwell.

Winter, Thomas N. 1999. "Roberto Busa, SJ, and the Invention of the Machine-Generated Concordance." *Classical Bulletin* 75 (1): 3–20.