Methods of quality, quality of methods. What does Roberto Busa have to communicate to digital humanists in the 21st century? From hermeneutics to performativity.

Marinella Testori <testorimarinella_at_gmail_dot_com>, King's College London

Abstract

Despite being also known as the "Father of Digital Humanities" owing to his pioneering contribution to the application of informatics to the whole ensemble of texts by the medieval philosopher and theologian Thomas Aquinas (1225-1274), the Jesuit Italian priest Roberto Busa (1913-2011) has still not been fully appreciated with regard to his development of a "hermeneutic informatics" [Busa 1999, 5]. Indeed this may represent a key concept to clarify what makes a difference between the common usage of computers in order to speed up procedures, and high-quality practices enhancing the role of informatics in shaping human interaction with machines. In other terms, Busa's interpretation of informatics may impact not only on the way in which digital resources and tools are developed, but also on the epistemological reflection about Digital Humanities. In this paper, by drawing from many of his texts, I will outline how this innovative "hermeneutics" is explained by Busa in terms of language dynamisms potentially leading to the development of what Johanna Drucker has described as a "humanistic-informed theory of the making of technology" [Drucker 2012, 87].

In his article titled "What is humanities computing and what is not?", John Unsworth suggests looking at humanities computing "as a practice of representation" [Unsworth], which he has articulated according to six propositions, among which are those concerning "humanities computing as a way of reasoning" and "humanities computing as shaped by the need for human communication". The basis for these suggestions outlined in 2002 may be found in the article "What is a Knowledge Representation?" by Randall Davis, Howard Shrobe and Peter Szolovits, which appeared around a decade before in *AI Magazine* and in the light of which Unsworth has developed his considerations: in this contribution, in fact, the authors have reasoned about the features of knowledge representation, which, among others, they have defined as "a set of ontological commitments", "a medium for efficient computation" and "a medium of human expression" [Davis, Shrobe, and Szolovits 1993].

It is worth noticing that, with their discussion, Davis and colleagues have anticipated the current debates regarding what digital humanities is and, more specifically, the boundary between a generic, widespread use of computers to do things, and computer applications adding value to academic research in humanistic disciplines because of their capability of representing human knowledge more effectively. Interestingly, in the second half of the twentieth century Roberto Busa (1913-2011), the groundbreaking developer of the *Index Thomisticus*[1], a linguistic corpus for Thomas Aquinas' texts, already proposed ideas with the potential to enlighten and steer the contemporary discussion in the field. Although Busa is known for having been one of the early pioneers to bring problems in language together with computing machinery, his contribution to the development of a "hermeneutic informatics"[2] is not so appreciated and recognised. There is the risk of pigeonholing his work under constraining labels or confining it only to specific sectors (for example, computing techniques applied to ancient languages). It appears to me significant, instead, that, since he worked on the development of a linguistic corpus, Busa established in this way a vital connection between humanities computing and the reflection on language.

In this article, I am going to outline the salient ideas put into practice by Busa by discussing some of his contributions as

a common thread. It is not a coincidence that in 1998 Busa admonished that "Informatics is already an ocean" [Busa 1999], so we need to be careful not to get lost. For this reason, I also propose many quotations from Busa's texts as a deliberate choice to let the reader know many worthy reflections provided by the scientist that are not so well-known so as to stimulate a debate. At the same time, I will be pointing out how my current experience of work on Busa's legacy is leading me to move ahead with the development of new perspectives requiring further consideration.

Methods of quality: Roberto Busa and the computing in the humanities

In his keynote speech *From Punched Cards to Treebanks: 60 Years of Computational Linguistics* at the Eighth International Workshop on Treebanks and Linguistic Theories (Milan, 2009), Busa sketched three main typologies of informatics currently in use: "documentaristic" informatics, comprising all the informatic services allowing efficient "information retrieval"; "editorial" informatics, referring to the wide range of "multimedia" devices for reading books, watching films, browsing the Internet; and "hermeneutic informatics", considering the "computerized text analysis, or language hermeneutics, i.e. interpretation, [...] of all our ways of questioning the *whys* of language" [Busa 1999, 5]. This latter was the informatic typology to which Busa devoted his inexhaustible attention during all his life. The focus on language, according to him, is strictly required by the nature of the relationships between man and computers, which interact by means of specific programming codes. Moreover, "the computer allows and exigently demands, as its specific capacities, an exhaustive, detailed, deep, quantitative knowledge, derived from huge amounts of natural texts" [Busa 1999, 6]. As a consequence, it is arguable that computers and human beings, technological devices and humanities are not competitors or antagonists, but they are potential allies, insofar as they both are "human expressions" [Busa 1999, 6]. Nevertheless, what do we really know about the language (our language) on which human communication is based and by which it is conveyed? Are we really aware of the intrinsic logic, the inner dynamism, and the psychological implications which are put in motion every time we communicate? Roberto Busa defined the "language that [is] unknown" [Busa 1999, 6][3], as signifying that establishing an interaction with the computer/machine implies a more profound level of language awareness than that to which we are accustomed. This raising of consciousness is necessary to introduce consolidated philological and linguistic methods to the "new qualitative dimensions" [Busa 1980] made available by informatics. In the light of this, applying computer methods to the humanities "can help us to be more humanistic than before" [Busa 1980, 89] because it leads us, first, to an inner journey through rational paths triggered by language expression. This conviction mirrors the current debate regarding the necessity of an epistemological reflection on the making of the digital humanities: for example, Stephen Ramsay and Geoffrey Rockwell have recently stated that "the understanding of underlying theoretical claims is the *sine qua non* of humanistic enquiry" [Ramsay and Rockwell 2012]. Roberto Busa would have certainly confirmed this vision and reasserted that these "underlying theoretical claims" primarily involve language dynamisms and linguistic issues. More specifically, these underlying implications concern the meaning, the "semantics" of words and sentences, which is not attainable by a mere quantitative production of a certain amount of data. Busa claimed that "we do not speak in words but in sentences. A sentence has a global meaning which is not the pure sum of the values of its single components. The heart of this problem is whether we are able to formalise the global meaning of sentences with something less than the whole sentence itself; in other words, whether we can succeed in identifying in each sentence something which can be taken as characteristic of its global meaning" [Busa 1980, 88].

Quality of methods: Roberto Busa and the hermeneutics of informatics

According to Busa's experience, a high-quality methodology in computing for the humanities should be focused on an accurate reflection on communication, because man and computer interact with the help of sophisticated languages which differ from common grammar and syntax and require constant development. Charles L. Isbell et al. have confirmed the language-oriented nature of the interaction established by human beings with the machine and suggested a reciprocal shaping-power of the language on the outside reality (the computer) and, *vice versa*, of the external world (the computer) to the language itself [Isbell et al. 2009][4]. A more conscious remark of this communicative essence may contribute to illuminating a controversial point in computing for the humanities: as Johanna Drucker has pointed out, "the challenge is to shift humanistic study from attention to the *effects* of technology (from readings of social media, games, narrative, personae, digital texts, images, environments), to a humanistic-informed theory of the *making* of

technology" [Drucker 2012]. With respect to this question, Busa's message is that the development of an "informed theory of the *making* of technology" cannot elude a renewed, unremitting consideration of language features and issues, which, in a kind of virtuous circle, are positively related to the *effects* of technology on our life. In this sense, the fruit sprouting from the "hermeneutic informatics" to which Busa referred in his last keynote speech *From Punched Cards to Treebanks* mentioned previously is what I propose to indicate as a *language-based hermeneutics of informatics*.

Moving ahead of Busa: from hermeneutics to performativity

This type of theoretical approach, along with the scepticism Busa expressed regarding possibilities for a machine to fully develop it [Busa 1990][5], clears an unexpected, irreplaceable space for the human role: in fact, although "the computer has even improved the quality of methods in philological analysis, because its brute physical rigidity demands full accuracy, full completeness, full systematicity" [Busa 1980, 88], investigating the meaning of a discourse, as well as the mind processes involved, requires that man instructs the machine to work at this level.

6

Busa found in the so-called *arbor Porphyrii* (the "Porphyrian tree", suggested by the ancient philosopher Porphyry to explain the Aristotelian *Categories*[6]), an effective way of representing how the human way of reasoning works. This is perfectly interpreted by the "trees" which constitute a treebank, a linguistic corpus characterised by a specific focus on the relationships between words and sentences. Busa was aware of the difference among the "Porphyrian trees", expressing "a graduated scalarity of similarities and differences" [Busa 2009][7] among genres and species, and the "treebanks", illustrating "relations of real and true dependency" [Busa 2009] from a linguistic point of view; nevertheless, he asserts that dependency-trees, provide "a syntax extracted inductively from computerised texts and workable by the computer according to its boundless capacities" [Busa 2009] that is a pedagogically powerful means for clarifying our interior logic and way of thinking and, consequently, of communicating[8].

7

Building a treebank for the syntactic and semantic annotation of the 11 million words constituting the *Index Thomisticus*, as it is currently being performed by the Research Group committed to Busa's legacy with the *Index Thomisticus Treebank* Project[9], is thus a method of quality *par excellence*, not only for the enhancement of linguistic properties in a text but also for casting light on some mental paths connected with language. My current experience as annotator of the IT-TB, in fact, has been increasingly persuading me about what I propose to define as the *performative nature of linguistic annotation.* While working in Busa's footsteps, I have been practically experiencing first-hand the insights he advanced in *From Punched Cards to Treebanks: 60 Years of Computational Linguistics* and, at the same time, I have been realizing how these insights can lead us ahead of what Busa himself could not have completely foreseen. The most intriguing fact of annotation of a treebank is that it enables the human annotator to deal with a living, branching and constantly evolving tree. Making a tree live, by establishing relations between words until it has reached its accomplished form, is the amazing vocation I have discovered, since, in the very act of annotating, I can make the text live. It really represents a kind of "making of language," to recall the title of the book by Mike Beaken, *The Making of Language* [Beaken 2011], in which is proposed an alternative view of looking at the origins of language, rooted not so much in biological features, rather than in cultural and technological developments. Treebanks, in my vision, represent one of these contemporary, worth exploring and pursuing developments, made available by the advancements of linguistics not only for modern, but also for no-longer spoken languages.

8

Conclusions

Language-based models of interpreting and building a language, along with the development of new technologies in the field of computing applied to the humanities, may represent theoretically and practically valid methods of contributing to the above-quoted challenges explained by Unsworth and Davis. Despite being software-embedded, in fact, these models and related technologies may be able to overcome the inherent limitations (e.g., strong object-orientation) of mere computational approaches not capturing the performative value of human acts of linguistic expression. In addition, owing to our language-shaped minds, they may offer a remedy to that "universal lament about the fragmentation of knowledge" [Busa 2009] that Busa ascertained, because "the human hunger for syntheses derived from microanalysis, continues to surge up, and not only in technologies, but also in linguistics, philosophy, psychology and theology" [Busa

9

2009] — in other words, in the context of general human knowledge.

Notes

[1] See the website http://www.corpusthomisticum.org/it/index.age

[2] This expression was used by Roberto Busa himself in his keynote speech, "From Punched Cards to Treebanks: 60 Years of Computational Linguistics, in *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories*, *4-5 December 2009, Milan*" [Busa 2009]. The full text of the intervention is not available in the Proceedings, but it can be acquired through the C.I.R.C.S.E., Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione, at the Catholic University of Milan (Italy), http://centridiricerca.unicatt.it/circse_index.html?rdeLocaleAttr=en.

[3] "Language that unknown" is a literal translation from Italian. The meaning is that, generally, we do not have a full, conscious awareness of the language in all its components.

[4] "The computing machine or artifact is typically manipulated through some language that provides a combination of symbolic representation of the features, objects, and states of interest as well as a visualization of transformations and interactions that can be directly compared and aligned with those in the world. The centrality of the machine makes computing models inherently executable or automatically manipulable and, in part, distinguishes computing from mathematics. Therefore, the computationalist acts as an intermediary between models, machines, and languages and prescribes objects, states, and processes".

[5] "Language is living, open and continually evolving. It is in tune with everything that is beautiful and new. Consequently, the epistemological methodologies of mathematical and physical sciences, which measure quantifiable physical entities, are not sufficient to dominate and grasp the logic of the signs we use to communicate knowledge".

[6] For a first, general information regarding the "Porphyrian Tree" see https://en.wikipedia.org/wiki/Porphyrian_tree . See also what is proposed about "Transcendentals and Predication" in [Goris and Aertsen 2013].

[7] The word "scalarity" appears in the original text, but the most proper term here would be "scale".

[8] "Dependency trees are very useful and very educative. They train us in an internal 'speleology' on our own logic, which in each of us is the spiritual centre of our own personal consistency and dignity. 'Knowing yourself' is a process that is never really exhausted".

[9] The website of the *Index Thomisticus Treebank* Project is http://itreebank.marginalia.it/.

Works Cited

Beaken 2011 Beaken, Mike. 2011. *The Making of Language*. Edinburgh: Dunedin Academic.

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

Busa 1990 Busa, Roberto. 1990. "Informatics and New Philology". *Computers and the Humanities* 24: 339-343.

Busa 1999 Busa, Roberto. 1999. "Picture a Man…Busa Award Lecture, Debrecen, Hungary, 6 July 1998". *Literary and Linguistic Computing* 14: 5-9.

Busa 2009 Busa, Roberto. "From Punched Cards to Treebanks: 60 Years of Computational Linguistics". In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories*, *4-5 December 2009, Milan*. Edited by Marco Passarotti, Adam Przepiórkowski, Savina Raynaud and Frank Van Eynde, http://convegni.unicatt.it/meetings_Proceedings_TLT8.pdf

Davis, Shrobe, and Szolovits 1993 Davis, Randall, Howard Shrobe, and Peter Szolovits. "What is a Knowledge Representation?". 1993. *AI Magazine* 14: 17-33, http://groups.csail.mit.edu/medg/ftp/psz/k-rep.html

Drucker 2012 Drucker, Johanna. "Humanistic Theory and Digital Scholarship". 2012. In *Debates in the Digital Humanities*, edited by Matthew K. Gold. Minneapolis/London: University of Minnesota Press, http://dhdebates.gc.cuny.edu/debates/text/34

Goris and Aertsen 2013 Goris, Wouter, and Aertsen, Jan. 2013. "Medieval Theories of Transcendentals". *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, http://plato.stanford.edu/entries/transcendentals-medieval/

Isbell et al. 2009 Isbell, Charles L., Lynn Andrea Stein, Robb Cutler, Jeffrey Forbes, Linda Fraser, John Impagliazzo, Viera Proulx, Steve Russ, Richard Thomas, and Yan Xu. 2009. "(Re)Defining Computing Curricula by (Re)Defining Computing". *SIGCSE Bulletin* 41, 4: 195–207, http://www.cc.gatech.edu/~isbell/papers/p195-isbell.pdf

Ramsay and Rockwell 2012 Ramsay, Stephen, and Geoffrey Rockwell. 2012. "Developing Things: Notes toward an Epistemology of Building in the Digital Humanities". In *Debates in the Digital Humanities*, edited by Matthew K. Gold. Minneapolis/London: University of Minnesota Press, http://dhdebates.gc.cuny.edu/debates/part/3

Unsworth Unsworth, John. *What is Humanities Computing and What is Not?*. http://computerphilologie.uni-muenchen.de/jg02/unsworth.html