THE WORK OF ROBERTO BUSA SJ: OPEN SPACES BETWEEN COMPUTATION AND HERMENEUTICS

GIANCARLO BOLOGNESI (†)
LUIGI DADDA
ADRIANO DE MAIO
TULLIO GREGORY

A review of the achievements of Fr. Busa over the course of his 60 years of work in the area of computational linguistics: internal hypertexts, the *systematization* of allographs, lemmatization, homographs and typologies; the *lexical system;* the laws of economy for graphemes, for semantic typology, for heterogeneity among terms, and of the two lexical hemispheres. Finally, the project of *disciplined languages* is mentioned, a response to the linguistic challenge resulting from informational globalization.

Keywords: Roberto Busa, Index Thomisticus, Computational Linguistics, Language.

1. INTRODUCTION

The aspect of the work of Fr. Busa which has most been emphasized in the communications media is that he was the first to use computers in processing words and texts, not merely numbers. This achievement of Fr. Busa would be remarkable even if we only took into consideration the quantity and dimension of his work over six decades of labor. For example, he analyzed and classified via computer some 11 million words in Latin, along with a similar quantity in twenty other languages: Albanian, Arabic,

Aramaic, Armenian, Bohemian, Catalan, Hebrew, Finnish, French, Gaelic, Georgian, Classical Greek, Old English, Italian, Nabatean, Portuguese, Russian, Spanish and German; and he did this in eight alphabets: Arabic, Armenian, Cyrillic, Hebrew, phonetic (IPA), Georgian, Gothic, Classical Greek and Latin. In addition, he has taken part in more than 100 international congresses on four continents, as well as having organized a congress in Tübingen (Germany) in 1960. He has founded two departments of computational linguistics, one at the Catholic University of Milan and the other at the Pontifical Gregorian University in Rome. And, during the last six years he has been invited by the Polytechnic of Milan to impart classes on philosophy and psychology in relation to artificial intelligence and robotics.

Nevertheless, the discoveries and conquests achieved through so much work have not always been equally celebrated, because many have been performed within obscure areas of linguistic research.

2. WHEN THE IDEA WAS BORN

1. The idea for automating linguistic analysis came to Fr. Busa during the years between 1942 to 1945, which were times of war, and for him, a time of preparation for teaching philosophy at the Pontifical Gregorian University. He is not arrogant about his discovery: "If the idea hasn't come to me at that time, it would have occurred to somebody else soon afterwards. *Causality* is nothing but Providence. At most, the merit comes afterward, because of perseverance."

3. Pioneering Methods and Terminology

2. He had to create both methods and terminologies. He could not seek them in bibliographies, nor in his readings, since these were entirely new ideas. Nevertheless, in the libraries of Rome, Milan, Munich, Paris, London and New York, he examined several hundred *concordances* in various languages. Upon noting what was required to produce them in Latin —and soon also in Greek and Hebrew— he drew from them precise methods and nomenclatures.

He allowed himself to be guided by the truth of things, following the counsel of Aquinas: "studium philosophiae non est ad hoc quod sciatur quid homines senserint, sed qualiter se habeat veritas rerum".[1]

4. "Hypertexts," even before the Word Existed

3. "*Quasi ab ipsa veritate coactus*",[2] from the beginning, and before current-day terminology existed (*hypertext*, SGML, HTML, TEI, XML...), Fr. Busa added three hundred distinct codes to each of the eleven million words —including *et* and *non*— contained in the corpus of Thomas Aquinas; these codes were encoded in 130 bytes, which specified many diverse values within the confines of morphology.

Now, at the end of his life, the project which he wishes to set in motion —the Bicultural Thomistic Lexicon, or LTB— will add to these previously entered codes, introducing others which will define the syntax of each word.

---

1. *In De caelo*, bk. 1 l. 22 n. 8. For St. Thomas, philosophy is the rational investigation of a universal synthesis of our living situation.

2. *Contra Gentiles*, bk. 1 ch. 43 n. 16, and in ten other places.

5. "SYSTEMATIZING" FIRST THE ALLOGRAPHS, THEN THE LEMMATIZATION, THE HOMOGRAPHS, AND FINALLY THE TYPOLOGIES

4. From the very beginning, the enormous size of the files forced Fr. Busa to *systematize* —a word that is very frequent in his writings— three textual situations: "allographs" —i.e. variants in the graphical form of a single word— , lemmatization and typologies of discourse.

In regard to allographs, he distinguished and reviewed the difference between those variants which were purely graphical, and those which were *formal* or stylistic.

5. Concerning lemmatization, Fr. Busa was one of the first to return to circulation the term "lemma".[3] This term is now included in dictionaries as signifying that first word form in a dictionary entry, which acts as a heading, representing the various inflected subforms and definitions. Fr. Busa systematized the procedures for lemmatization, distinguishing clearly between that which was only morphological and that which was syntactic.

Morphological lemmatization, which is applied to the various forms of a word as it occurs in various contexts —in the Thomistic corpus, there are 150,000 different word forms— is organized in a *tripartite* manner (invariable words, declinable words, and conjugatable words), and turned out to be the most practicable for the computerization of texts of large size.

Syntactic lemmatization was later applied later to the 11 million context sentences, one by one, classifying each word according to eighty aspects of speech.

Fr. Busa has always been skeptical about automatic lemmatization, but is not against a semi-automatic process, once the first important part has been done by hand. Nevertheless, he recognizes that the first has the methodological value of dealing

3 The Greek word *lemma* entered Latin only in the postclassical period, and remains today embedded in terms such as *dilemma*.

synchronically with the formalization of the structures present at the surface of our expression. In fact, it is enormously important to him to distinguish, for example, between the two systems of interior forces, i.e. understanding and expression.

6. The process of lemmatization forced him to immediately face the linguistic phenomenon of *homography*, which was never systematized prior to the advent of the computer. Indeed, we all speak and read by phrases, which nearly always prevent this homography from being noticed. Fr. Busa began to study it by means of its causes, types and causality: he discovered (even without mentioning those which occur between parts of speech, or between words of diverse languages) that at least half of the eleven million words in the Latin texts of Aquinas turned out to be homographs for one reason or another. Thus, he had to individuate all the homographic *forms* within well defined limits, to discover how to evaluate their probability of occurrence in the Thomistic corpus, how to create a repertory of all of them —insofar as they were *possible*, at least— in order later to be able to distinguish the most important ones, leaving the rest for the future. In fact, he made this differentiation for 600,000 contexts. The necessity of such systematization is obvious for the validity of any computerized elaboration of texts, given that the computer can only work on the physical form of the signifying signs.

7. The typologies of discourse were discovered to be numerous in the literary genres of the sampling of works analyzed in 20 different languages: scientific abstracts, works of theater, letters, literature, manuscript editions...

In the *Index Thomisticus*, Fr. Busa marked each term with at least two of the following contextual codes: 1. the author's own discourse; 2. a literal quotation; 3. a quotation *ad sensum*; 4. a brief sample of initial words (*incipit*); 5. a reference to another text; 6. a reference to the current text itself; 7. in a digression; 8. its weight in the flow of the discourse, whether central or peripheral.

8. These careful and prolonged preparations, and especially the lemmatization, obtained their recompense through permitting and accelerating other more advanced research.

6. THE FIRST "LEXICOLOGICAL SYSTEM" OF AN AUTHOR

The *lexicological system* of the Thomistic corpus is the first and, even today, the only existing such system, if we understand such a system as the final result of analysis and later synthesis of a closed linguistic universe, according to all of its elements of morphology, syntax and lexicon. This is a new kind of linguistic document: an integral quantitative and statistical classification of the main, most important and fundamental expressive elements of a linguistic system.

The 294 pages of the *Treatise on Lexicology* by Fr. Busa[4] provides a summary of a general system and three subsystems (homography, typology and quantity) of the forty tables of the 9th and 10th volumes of the *Index Thomisticus*,[5] which summarize in 2,470 pages the detailed data found in the 8,022 pages of the previous eight volumen.[6]

Understood in this way, the *lexicological system*, thanks to the computer, has initiated a new discipline, if not in name, then *de facto*. Indeed, a *lexicology* understood in this way would correspond to a *computational linguistics* considered in the full sense of its final objective, i. e. to provide integral, classified and

4. R. BUSA, *Il libro dei metodi, t. 6: Trattato di Lessicologia*, CAEL, Gallarate, 2001, 264 pp.

5. R. BUSA, *Index Thomisticus. Sancti Thomae Aquinatis operum omnium indices et concordantiae, vol. 1: Sectio prima. Indices, t. 9: Systemata lexici, I: Systema lexicologicum: Tabula 1: Systema lemmatum. Tabula 2: Systema formarum A-O* (Frommann-Holzboog, Stuttgart, 1980) XVI, 1257 pp.; *Ibidem, t. 10: Systemata lexici, I: Systema lexicologicum: Tabulae 2 (finis)-5. II: Systema homographiae: Tabulae 6-12. III: Systema typologicum: Tabulae 13-26. IV: Systema quantitatum: Tabulae 27-38* (Frommann-Holzboog, Stuttgart, 1980) XII, 1210 pp.

6. *Ibidem, vol. 2: Sectio secunda. Concordantiae operum thomisticorum. Concordantia altera, t. 1-8* (Frommann-Holzboog, Stuttgart, 1979) XVIII+1286, 1282, 1286, 1293, 1287, 1300, 1270, 1297 pp. In fact, he had already calculated the total number and percentages of the categories of 11 million words, first on 150,000 word forms and later on 20,000 synthetic lemmas (each one corresponding, on average, to four in Latin dictionaries of usage).

statistical linguistic syntheses obtained from an ever-growing number of texts —that is, from closed linguistic universes— as a documentary base. It is evident that such a lexicology would contribute greatly to a healthy methodology for scientific research, including for the *human science* of linguistics.

7. THE DISCOVERY OF FOUR LAWS (OR ALMOST)

10. On the basis of this lexicological system, and with thousands of man-hours of teamwork, Fr. Busa was able to work out on his own[7] four *discoveries*, in the etymological sense of the term: something knowable which only now has been brought into the light, an invention in the sense of an encounter, passage from that which was implicit but hidden to that which is explicitly known.

11. The first discovery was a type of law of economy in the relation between the number of words and that of the various chains of characters that comprise them. Specifically, he divided each word —that is, each lemma, after separating it from its declension morphemes— dividing that which was constant into a maximum of three segments (not morphemes!): initial, central and final. He applied the name *string* to each of the equal sequences of strings which were found in different words, combined with other strings and ignoring their meanings.

It turned out that 1,500 chains of characters (which were later able to be reduced) between 1 and 12 letters, combined together, were able to produce all the 11 million words (save for 4,000, which were identified) of the entire Latin corpus that was analyzed. This is the documented fact, although it can be supposed that it would be valid also for other languages, at least of analogous

7. During his long life, Fr. Busa has noted how, since the beginning of time, *that which is* new spreads slowly, due to the frictions and obstacles set up by established knowledge.

type. We do not know whether this pattern has been tested for and proved to exist in other languages. In any case, it would be interesting for the compression and electronic transmission of texts.

12. Fr. Busa distinguished the registry of the heterogeneity of the words from the registry of their *semantic type*, understanding the latter as a relation between sign and knowledge (*signifier-signified*) within a bidirectional operative arc, from knowledge to expression and vice versa.

13. The following is a schematic summary of these semantic types, omitting the decimal codes:

> 1% are explicit deictic words (distinct from those which are always implicit in the declensions of the 1st and 2nd person singular and plural of Latin terms) which are a part of the personal pronouns and of the demonstrative pronouns and adverbs. They do not express mental images, but rather *knowledge* about a presence (whatever it might be).
>
> 2% are proper names. These are word-labels which signify a one singular individual at a time, although at times also can signify collectives.
>
> 6% are common nouns, which denominate specific types of *objects* or *things*. For example, plant, horse, car, sandwich...
>
> 46% are those adjectives and verbs which specify the *aspects* of the things or objects: activity, passivity, quality, dimensions, figures, smells, flavors...
>
> 35% are particles, prepositions, conjunctions... which signify direction, relation, correlation...
>
> 8% are *vicarious* words which point to other words, concepts or things. They are pronouns or pronominals (in Latin there are no articles).
>
> 1% are words which signify persons or intelligences *beyond the physical*, which we can call *invisible*.

14. Many will remember the analogies or correspondences between this categorization and those of the supreme categories of reality from Aristotle and Kant.

15. Basing himself on this classification, Fr. Busa has extracted two consequences: first, that in any lexicon the words are heterogeneous. And this is true to the point that he attributes the meager results of the statistics concerning word frequencies in natural texts to the fact that words are normally counted as if they were homogeneous, like numbers within the same calculation. For each of the seven groups noted above, one should perform a separate calculation of statistics, and only later join the distinct results into a superior statistical result.

16. The second consequence, discovery or rediscovery, was that in every lexicon there can be found two hemispheres. One, which expresses the internal logic of the discourse, consists of few, normally brief words, which are repeated frequently and are equally present in all types of discourse. Sometimes these are called *grammatical terms* or *function words*. The second hemisphere, which specifies the message to be communicated, consists of many diverse words, frequently long, which vary according to the content of the discourse (also called *content words*), and whose frequencies are always inferior to those of the first hemisphere.

In the case of St. Thomas, the deictic words, relational words and vicarious words add up to 44% of the total corpus. Proper names, aspect terms and invisible objects make up the remaining 56%. In addition, various adjectives and universal verbs of high frequency should be attributed to the first hemisphere. In fact, ordering the 150,000 distinct word forms in the *Index Thomisticus* by their frequency, we discover that the 80 most frequent words make up 41% of the corpus, and the 800 most frequent make up 68%.

17. Fr. Busa believes that substantial progress is to be hoped for in the domain of computational linguistics through the employment of all the information mentioned up to now.

8. The Linguistic Challenge of Automatic Translation

18. Between the years from 1950 to 1965, Fr. Busa played an active role in the effort to develop technologies for automatic translation, research which was sustained by financing from the Pentagon. This economic support stopped suddenly in 1965, because the linguistic sciences were not providing precise data for a computer program that would translate texts to another language. Forty years later, substantially the same challenge has arisen, with other names and other motivating factors, due to the globalization of communication networks.

9. The Perspective of "Discipline Languages": A Proposal to the European Union

19. During an official linguistic congress of the European Union held in Strasbourg in 2002, Fr. Busa formulated a strategic proposal, which he called *discipline languages*; this concept was the fruit of his prior research in the profundities of Latin expression, and of all that which he saw and lived during sixty years working in computational linguistics.

20. Several decades ago, Fr. Busa had emphasized the fragmentary nature of the work he had performed via several vivid expressions —in the style of heroically audacious commands during a battle— concerning his focus on literary texts, three of which are included here:

- "A mile of algorithms built on top of an inch of text,"
- "Only the second floor, without the first,"
- "Ten people building the first mile of a highway, through the same forest and in the same direction, without anybody building the second, third or fourth mile, etc."

21. In Strasbourg, Fr. Busa wondered to himself, and inquired of the other attendees, whether the following would be audacity or utopian thinking (schematically summarized here):

• A community initiative, globalized and synchronized, in three phases:

First phase:

• That, in every principal language,

• based on university textbooks in each of the principal academic disciplines, transcribed in electronic format,

• the *lexicological system* would be extracted —in the sense described by Fr. Busa— for each of the selected academic disciplines,

• in order to combine them later into a single system for each language which would specify and quantify the convergences and divergences in lexicon, morphology, and syntax.

Second stage:

• At the same time, the systems for each individual language would be merged into a single *interlingual* lexicological system which would contain, in computer format, the geographic map of the correlations of convergence and divergence. This would be a detailed repository of "discipline-specific languages," with percentages and links between the correspondences and between the divergences in the lexicon, morphology and syntax of each language with regard to the others.

Finally,

• In each language, a manual of the discipline language would be defined and published, with lexicon, morphology and syntax, in order for it to be employed as the *input* for network messages.

• At the output end, the message recipient would be able to request from the central server, a translation to a target language.

Each one of these three stages would produce research documents and synthesizing conclusions based on factual, publishable and useful data for linguistic research purposes.

Giancarlo Bolognesi (†),
Università Cattolica del Sacro Cuore, Milano

Luigi Dadda, Politecnico di Milano

Adriano De Maio,
Libera Università degli Studi Sociali, Roma

Tullio Gregory, Università di Roma “La Sapienza”