Paper by Father Busa
Milano – 5 December 2009
TLT8 Workshop

From Punched Cards to Treebanks.
60 Years of Computational Linguistics

001
I was ordained priest on the last day of May 1940, my ordination was brought forward because Italy was about to enter the War; which duly happened 10 days later. We were twenty Jesuits, all slated to be military chaplains. I had been assigned to Bordeaux in France, where a detachment of Italian submariners was serving alongside a German one. The Provincial of the Jesuits called me in and asked me: "Would you like to be a teacher?". I, being enamoured of my chaplain's role, and also because, when one is young, one has a magnificent unconscious spirit of adventure, said no. Then he smiled broadly from ear to ear, and said to me: "Oh well, Father, that's what you are going to do". So – to put it crudely – they packed me off to the Gregorian University in Rome for the duration of the War, to do a university teaching qualification in philosophy on Saint Thomas.
In the autumn of 1941, Father René Arnoux, a French Jesuit who was a specialist in Plotinus, and at that time very well known, assigned me as subject for my PhD thesis the doctrine of presence in Aquinas.
I worked at my thesis over the disturbed period of the War years and I defended it in Rome shortly after the Epiphany in 1946.
The first 6 months I spent on it had led me to the conclusion that the Thomistic doctrine of presence was not expressed by Saint Thomas by the word *praesentia-ae*, but by the preposition *in*, which Aquinas used in working on the eleven values schematized by Aristotle; and it had to be so, because *praesentia* is not a simple or unitary notion, for it is the expression of a system, that is a system of reciprocal experiential acquaintances between A and B, and B and A.
It ramified as it were orthogonally: *unius ad alterum* (through efficacy) and *plurium ad unum* (produced by one and the same).
In was defined on 2 levels: locative (or collocative) spatial presence (or co-presence) and generative presence in the author-work relation, origin of opposite relations – and it is a simple primary elementary notion, which however enters into the pregnancy of the syntagma *causa propria*, where it involves the mental generation of the idea or formula of what is produced.
The bigram *in*, when in composition, that is as a prefix, has 2 opposite or contradictory values: one negative, as in *impossibilis-e*, and one positive, as in *internus-a-um*; but it is only a positive preposition when not a prefix but standing alone.

002
I now know that *in* occurs both as form and as lemma a total of 249,459 times in the writings of Aquinas alone. But in those days

I simply assumed that the (then unknown) number in question must be so enormous that it would not be countable. I therefore limited myself to writing by hand around 10,000 cards, each of them containing a 'characteristic' expression using *in*.
I once met a gentleman on a flight to Manchester in England who asked me: “Father, would it not be much more interesting to conduct research into the word *inn*? (with two *n*s - which in English means ‘hotel’ or ‘pub’ - one of our earthly pleasures ...)”.
I played a great number of games of solitaire with these cards, and the result was a book that was my PhD thesis, published in 1951 with the title "La terminologia tomistica dell'interiorità - Saggi di metodo per una interpretazione della metafisica della presenza" (*The Thomistic terminology of interiority - Essays on a method for an interpretation of the metaphysics of presence*) .
Two assured results emerged from this: the first was that the smallest elements of discourse are pregnant with basic because elementary philosophical significance. The second was that I would render a useful service to culture if I were to produce a concordance of all the individual words of Aquinas (*et* and *non* included and without exclusion of any hapax) and if moreover I were to classify them not only according to the morphology of lemma and form, setting out the vast number of their homographs, but also according to the most (if I can put it like this) speleological genetic classification of each lexicon, ie (my own approach) according to types of semanticity, or of the different types of relation between sign and signified or concept:
1) deictic words (plus proper nouns): e.g. *ego, tu, hic, ille*;
2) things or objects (*substantiae*, *personae*): e.g. *cibus, mons*;
3) their aspects: e.g. *accidentia, dimensiones (parvus), qualitates activo-passivae (auctor, opus)*;
4) correlating words: e.g. *est* (when a copula), *propter*;
5) vicarious words: e.g. *quis/quid*, *tertius, quartus* etc.;
6) names of invisible persons: e.g. *angelus*, *deus*.

003
In all of the world of culture the increasing availability of computers since the fifties has intensified the requirement for scientific rigour - that every serious hermeneutic be backed up by a numerical and statistical account of its graphical structures, making it possible for all users to check the whole of it.
So it was that I began to look for machines that would do linguistic analysis, or “lyophilize” (‘freeze-dry’) a text, word by word. I had the opportunity of going to New York to see IBM and present my request in writing: I asked IBM if they had machines to do this work. Mr. Thomas Watson sr, the old lion who founded IBM, presented me with a report by one of their technicians who thought it impossible to make machines do what I wanted them to do and asserted that I was more American than they were: it was a euphemism for telling me that I was a bit 'nutty'. Some saint helped me to reply: “Mr Watson, do you think it right to say that a thing is impossible, even though nobody has even

tried to do it?". Then I got out a card I had found in the lobby – at the time, in IBM, there was a passion for slogans, they were all over the place(remember the famous: "Think!") – on which was inscribed: "The difficult we do right away, the impossible takes a little longer". It was not very modest, but at that moment it was what I had to hand. I made as if to return the card, then Mr Watson looked at me and said: "All right then, Father, I will give you what you need to test it out. If you succeed, IBM will help you to bring your project to a conclusion. But on one condition: that you do not change the name from International Business Machines to International Busa Machines". I took this as a humorous remark, but I noticed that from that moment, wherever I went in IBM, I was given the red carpet treatment: they thought I was in "cahoots" with the great chief. Without making any written commitment, IBM backed me for thirty years – which is the time it took to complete the first work on IT, with 9 million words by Aquinas and 2 million words by other Latin authors. But in certain ways, it is still way out ahead today. I continued to work on millions of other words in 18 different languages and in 8 different alphabets. This work was necessary in order to be able to test drive the method and create the terminology and procedures. Also to find the funding. Anyway, it went well.

004
To show how unforeseeability is built into life – for life is not Cartesian – suffice it to say that at that time the first goal of the programme was to create a card-index of 12 million punched cards, enough to fill a series of cabinets 90 metres long, 1.20 metres high, one metre deep. The overall weight amounted to around 500 tons. On the back of each card, in the spaces between the unperforated lines, we had stamped a box of 11 lines. It was only when I had already used 6 million punched cards that Providence came to my aid with the invention of magnetic tapes. Just imagine that in demonstrations in the days of punched cards, the representatives of IBM lauded the 'dizzy' speed of the automation, by dint of which, with the electronic sorting machine, 20,000 words could be put into alphabetical order in one hour! This may seem laughable today, but in those days it really did seem impressive.

005
The second stage was the magnetic tapes. I put 1,800 of them to work. They were quite a handful to work with, especially when it came to the connections between the sections. Once they were all connected up, the resulting tape stretched to about 1,500 Km. After this work on the magnetic tapes was completed, I condensed the whole into twenty of those big pizza-style tapes. From these, with a series of programmes, were taken the 56 volumes of IT (about 70,000 pages), and later the cd-rom.

006
The cd-roms, the third stage, still in use today, have a capacity of about 700 mb. On my tapes I had about one billion, 630 million bytes. We had compressed them by certain mathematical methods into 200 million bytes, so that – too much grace, Saint Anthony! – I still have space to be filled. Such is the story.

007
The present is the Internet. The *Index Thomisticus* can be accessed at the site www.corpusthomisticum.org.
Moreover the syntactically annotated data of the *Index Thomisticus* Treebank, about which I will say a few words, is also accessible.

008
Today the informatics handling of written texts is split between different disciplines which are generally confused by lay persons. I will reduce them to three.

First: the most developed is the one that I call *documentaristics*.
Fifty years ago it was called scientific documentation. For many years, from 1949, in fact, I found myself by the nature of things linked to the "Deutsche Gesellschaft fuer Dokumentation" and the "American Documentation Society". Then it got called "information retrieval", and today "information science".
This is focussed on data banks, from which the desired information is recovered at the time it is required.
Although it is the pure applicative result of scientific studies, this is more a form of service than scientific research properly so-called.
It is in fact becoming an infrastructure for social communications, like all the networks for the transmission of energy or the transportation of things and persons.
Hypertextuality emerged within this, but today it is moving away from it towards areas of genuine research into that which is not yet known.

The second type of informatics is editorial, today it is developing with CD-ROMs, including multimedia, and with the internet.
Multimedia CDs are a new form of book, for audiovisual reading, which is reducing the distance between presential oral expression and the written message.

The third type, which is (so to speak) my own, we may call hermeneutic informatics.
This sets out to deprogramme what is at the source of a written text.
It is a humanistic research, if humanism is still the study of the expression of man.

It is important to distinguish clearly between these three types of informatics. Research activities are all too prone to result in overlap, but this tendency needs to be contained within reasonable limits, so that in the processes of conquest of knowledge the production of valid construction blocks can go forward in the midst of the all too numerous sterile repetitions.

009
I have been following the specialist journals for years, and I see that most of the research amounts to a kilometre of algorithms based on a centimetre of foundation, and sometimes it even reduces to little more than research "miniatures".

010
By contrast my own practice has been to start from a kilometre-long base and rise by a centimetre all the length of that kilometre; then by a second centimetre all the length of that kilometre, and then by a third centimetre all the length of the kilometre, and so on.
Where the study of language is concerned, viable conclusions are obtained from complete classifications of large quantities.
Samples in textual researches are valid if they are limited to signs, as distributed physical entities, but when the semantics of them is studied, that is if their signifieds are analysed, only very rarely can general laws be deduced from them: just as an analysis of the musical modules and stylistic features of Wagner, would not tell me about those of Rossini or of Strauss.

011
Let me indicate by an analogy a second problem, arising from the nature of human life. A hundred companies set about constructing each on its own account a motorway in the forest. After a year, we will have one hundred first kilometres, but no second kilometre, no third and so forth.
The general lack of coordination leads to repetitions and overlaps, it also neutralises the economies that could be generated by the adoption of universal agreed standards.

012
Distant though they may be, the ultimate aims of my "Lessico Tomistico Biculturale" (Bicultural Thomistic Lexicon) (LTB) today are twofold.

The first aim is lexical: to research the entire word list of the IT texts, in order to produce a compendium dictionary with conceptual translations which tell us for each Latin word of the *corpus thomisticum* which words in many of today's languages best express the various concepts that this Latin expressed then.

Further, to verify and document which of the IT lemmas should be subdivided, on the ground that in different contexts they present with fundamentally different meanings.

The second aim is to construct a summa of the entire syntax of Aquinas with statistics and percentages of each grammatical element, including punctuation marks (this is the IT-Treebank project): this will then serve as a metre or measure to compare or contrast the Latin grammar of St Thomas with that of others in other languages as well. Just as is being done elsewhere for classical Latin, for example in Boston, and for Biblical Latin in Oslo.
IT has added to our 11 million Latin words 300 bytes of internal hypertexts, alternating in 130 positions, which defined the morphology (grammatical analysis) of each word in each of its expressions.
The LTB is beginning now to add to these the internal hypertexts of the syntax (logical analysis,) and it is drawing on them to rectify all the IT homograph morphocodes, so that IT + LTB together become a complete and perfect record.

013
A first syntactic analysis will be applied step by step to the electronic IT text, and then it will be transferred to the "*corpusthomisticum*" website.
This constitutes the first task and the particular contribution of LTB:
a) rectifying the morphocodes and the lemmatisation;
b) distinguishing 3 levels of paternity: St+, St-, Aa.

014
At the same time and in parallel, the structuration of the IT in syntactic trees with dependencies (IT Treebank) is also being initated.
Trees are a further type and manner of expression, or a system of signs which expresses, describes, visualises, communicates the quantity of the relations or correlations existing between words.
In the series of words of a text, the trees of syntactical dependencies do not codify the morphology of the words, but, when this is known, they do give their 'why' so to speak, as the skeleton of the whole: they show that it is those logical trabeations which compose the expression in unity; and therefore they correspond in their way to a multidimensional representation of the complex reality of things or notions.

015
There is a real sense in which syntactic trees possess an ancient pedigree.
In fact Porphyry (known as "the one from Tyre", 233-304 AD) born in some Middle Eastern country - we are not quite sure which - and

eventually a disciple of Plotinus in Rome, wrote the *Isagoge* (introduction) to the *Categories* of Aristotle.
Down the centuries this work, translated by Boethius, became the textbook of Scholastic logic.
The “praedicamenta” (mental notions, common and universal, recognised-in and predicated-of objects, or of things or substances), are disposed in a “tree” that has become famous as the *arbor Porphyriana.*

016
In the *arbor Porphyriana* the Aristotelian categories, characterising cosmic realities, are however disposed according to similarities of contents, descending from the thinnest, most generic, and most universal, towards those progressively more enriched by specificative differences; they are telescopic in the sense that each one that follows includes the preceding one.
It does not therefore express relations of real and true dependency, but only a graduated scalarity of similarities and differences denominated as much by the types of their interaction as by the types of the areas of their field of action.

017
Dependency trees are not of the same species as the Porphyrian.
Nor are they of the type of taxonomies or genealogies, though representable as a tree.
In fact they present only the logical network that synthesises in unity all the words within the segmentation in expressions of every discourse.

018
Dependency trees are very useful and very educative.
They train us in 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.
Overall the personnel working at “computational linguistics” are starting to put together, piece by piece, collegially, testing and retesting, each in his own language, a syntax extracted inductively from computerised texts and workable by the computer according to its boundless capacities.
The ultimate goal is a codification (both of logical dependencies and of the different lexicons) that becomes “universalised”, so as to be interchangeable among the various languages.

019
Dependency trees are the stones of a path that ascends towards unifying syntheses.
Today’s universal lament about the fragmentation of knowledge, which slides so easily into its disintegration, shows that 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.

020
These microanalytical representations will then be applied as further internal hypertexts of syntax to the 11 million Latin words on IT, in addition to those that indicate there both the morphology of each word and some basic typologies of each expression, these microanalytical representations will therefore be applied as further internal hypertexts of syntax.

021
The *Corpus Thomisticum* has the advantage of being a linguistic universe that is not only gigantic but also closed and immutable, in which all the linguistic situations can be counted right down to the last unit of the last comma.
Moreover the work of Aquinas sums up the 40 centuries of Mediterranean civilisation which preceded it: from the Mesopotamian of the Bible to the Greek, Roman, Christian and Arab.
And all this with a wonderful mastery of logical coherence, lexical correctness, and regularity of discourse.
It is not for nothing that the fact of being an encyclopedia of human life – though in an era when history, life, and science had not yet developed as they have now today – and even more the fact that it is not an encyclopedia parcelled out according to the alphabetical order of the words, but rather already digested and organised into a cosmic synthesis where each particular invokes the whole, have placed St Thomas among the elect perennial mentors of humanity, along with Homer, Plato, Aristotle, Augustine, and Dante – to name only the period up to 1300.

022
"Moriturus vos salutat"...!
This hoary and trembling little old fellow looks with great affection at the young people present here: a field of wheat that is turning golden and that promises abundant future harvests.
I want my spiritual testament for you only to be a testimony of lived experience: it is very beautiful, deeply intellectual, and scientific to spend one's own life – as well as continuing it in others – in working intensively and methodically to continue the human work towards the 'more and better' than before in the name of God.
Soon the Lord Jesus and His Mother will take me to Their home.
From there I promise you my orbiting presence. It is actually the eternal universe of the Spirits, of the Intelligences and of the souls which contains and upholds in its arms this sensible world of transient temporality, and it will swallow us up at our last dive.

But I also promise faithfully that I will not appear to you at night as an ogre with those punched cards in my hand in which the dampness of the inks made it maddeningly difficult for me to see where to calibrate the print lines between the perforation lines, so that these would never break up.