Andrea Colli

Making computers *noble.* An experiment in automatic analysis of medieval texts[1]

1. Introduction

Computer analysis of texts, creation of databases[2], hypertexts and digital editions are not the final frontier of research anymore. Quite the contrary, from many years *they* have been representing a significant contribution to medieval studies. Therefore, we do not mean to make the computer able to grasp the meaning of human language and penetrate its secrets, but rather we aim at improving their tools, so that they will become an even more efficient equipment employed in research activities[3].
On a closer view, giving serious consideration to this progress of the matter signifies keeping an eye – from an atypical viewpoint – on the question of the intricate relationship between philosophy and philology regarding the medieval studies. Now, that working for a critical edition of a philosophical texts does not imply a theoretical effort and, by contrast, that a philosopher tackles the speculative problems without taking into adequate consideration the textual basis, is obviously a stereotype. In fact, in most cases, the two perspectives of investigation are not empirically separable: a good editor must be a good philosopher and a primary knowledge of paleography should be requested to be a good scholar in medieval studies. However, an aspect, by which these two approaches could be discriminated, resides precisely in the different use of digital and automated tools. Elaborating critical editions of manuscripts or identifying the principal sources of a medieval text implies usually a massive use of databases, open libraries of digital archives. Actually, in order to classify recurring quotations, well-founded information can be obtained also by using the most common and accessible search-engines.

On the contrary, for a theoretical lecture of doctrines, a detailed examination of textual data seems frequently to represent only a sort of premise (sometimes a pretest) to philosophize, therefore computer analysis on the texts does not play a so crucial function.

A peculiar trait of this methodological disparity is related to the insufficient aptitude, by several medieval scholars, to take advantage of all potentialities offered by actual tools for automatic text processing. In other terms, several computer technologies, conceived for analyzing different kinds of texts, are – in my opinion – not yet adequately employed in the medieval studies, although the use of a common language (Latin) as well as the presence of stereotypical and recurrently adages or quotation should represent an advantage.

This paper is thought as a sort of technical report with the proposed task to verify if an automatic identification of some word associations within a selected groups of medieval writings produces suggestions on the subject of the processed texts, able to be used in a theoretical inquiry. In other terms, by adopting digital technologies as an unitary perspective, the goal is to explore in which terms it is possible to progress from a lexical or linguistic approach towards a philosophical one without methodological alterations or interruptions. Could computer analysis provide continuity between philology and philosophy?

The remainder of this article is organized as follows. Section 2 focuses on the task description, by giving a short overview of the inventory of the data and by defining the KWIC adopted for processing them. In the section 3 some algorithms, accessible in a suite devoted to textual analysis, are tested on the selected data. The results of the task are discussed in Section 4, while section 5 concludes this paper.

2. Task description

2.1. Sources

The sources of data consist in Albert the Great's *Opera omnia* (37 treatises, considering only the texts edited within *Editio Coloniensis*)[4] and Bonaventures's *Opera omnia* (12 treatises including in Quaracchi's edition)[5]. Both series of writings are available in a digital version by

Alberti Magni Opera omnia (Aschendorff)[6] and *Latin Library Series A* and *B* (Brepolis)[7]. Albert's texts were defined «cluster *Alb*», while Bonaventure's works «cluster *Bon*».

2.2. KWIC

As described in the section 1, the two defined clusters of texts was processed with the intention to outline the semantic function assumed by some keywords: i.e. the noun *nobilitas, nobilitatis* and the adjective *nobilis,e*.

2.3. Data

All the occurrences of *nobilis* and *nobilitas*, included in clusters A and B, were extracted with three lines of context[8]. In so doing, from the two original clusters, we circumscribed two selected group of texts, «sub-clusters *Alb1*» and «sub-cluster» *Bon1*, characterized by the following properties.

Table 1: Source data

	Alb1	*Bon1*
texts	35	11
words[9]	24764	31141

In order to process the extracted texts by tools for automatic analysis, words were lemmatized by *TreeTagger*[10], as displayed in the following example:

Table 2: Example of outputs after lemmatization with *TreeTagger*

Input	*Grammatic Annotation*	*Output*
Quaeritur	V:IND	quaero
etiam	CC	etiam
in	PREP	in
quo	REL	qui

illorum	DIMOS	ille
verius	ADJ:COM	verus
et	CC	et
nobilius	ADJ:COM	nobilis
sit	ESSE:SUB	sum
et	CC	et
evidentius	ADV	evidenter
secundum	PREP	secundum
esse	ESSE:INF	sum

During the process, *TreeTagger* did not recognize some words: in *Alb1* the «unknown» lemmas were 1731 (5,88%), while in *Bon1* 640 (1,75%). Regarding *Alb1*, more unfailing output was achieved by a supplementary lemmatization circumscribed only to the «unknown» words (973; 3,93%).

Both lemmatized groups of texts were imported in a database structured as follows:

Table 3: Example of database including lemmatized corpora

Group	*Work*	*Input*	*Grammatic Annotation*	*Output*
Alb1	De V universalibus	Quo	N:abl	qui
Alb1	De V universalibus	Enim	ADV	enim
Alb1	De V universalibus	nobilius	ADJ:COM	nobilis
Alb1	De V universalibus	Et	CC	et
...	...	...	...	...
Bon1	Breviloquium	Et	CC	et
Bon1	Breviloquium	Sic	ADV	sic
Bon1	Breviloquium	describit	V:IND	describo
Bon1	Breviloquium	totum	ADJ	totus
...	...	...	...	...

The following lemmas were excluded from our analysis:

- conjunctions (CC);
- dimostrative adjectives and pronouns (DIMOS);
- relative adjectives and pronouns (REL);
- numeral adjectives and pronouns (NUM);
- determinative adjectives and pronouns (DET);
- possessive adjectives and pronouns (POS);
- adverbs (ADV);
- prepositions (PREP);
- abbreviations (ABBR);
- the remaining «unknown» words.

The terms were extracted from column *Output* in order to create two new *corpora Alb1lem* and *Bon1lem* including respectively 15740 and 16454 words.

Both *corpora* were imported in T-LAB, a software framework including linguistic and statistical tools for content analysis and text mining[11]. During this phase *Alb1lem* and *Bon1lem* were segmented into elementary contexts (EC)[12] in order to help user exploration and, above all, to make analyses that require the co-occurrences[13] computation.

Each tool for textual analysis was applied separately to *Alb1lem* and to *Bon1lem*. An examination of differences and similarities between the outputs of the two groups of texts is presented in section 4 of the paper.

3. Tests

3.1. Word associations to *nobilis* and *nobilitas*

Co-occurrences that most recurrently determine the local meaning of our KWIC (*nobilis* and *nobilitas*) were picked up. Cosine is the used association index (or similarity coefficient)[14].

The output of this process is shown in the following[15]:

Table 4: ranking of word associations to *nobilis* in *Alb1lem*

	LEMMA B	*EC(AB)*	*PROXIMITY*
1	*sum*	281	0,989
2	*nobilitas*	127	0,66
3	*habeo*	119	0,643
4	*anima*	115	0,632
5	*natura*	95	0,576
6	*corpus*	91	0,564
7	*causa*	78	0,522
8	*bonus*	71	0,497
9	*possum*	65	0,476
10	*intelligentia*	62	0,468

Table 5: ranking of word associations to *nobilitas* in *Alb1lem*

	LEMMA B	*EC(AB)*	*PROXIMITY*
1	*sum*	130	0,673
2	*nobilis*	127	0,66
3	*habeo*	54	0,429
4	*natura*	45	0,401
5	*bonus*	38	0,391
7	*corpus*	36	0,328
8	*anima*	36	0,291
6	*causa*	34	0,334
9	*possum*	32	0,344
10	*ratio*	29	0,336

Table 6: ranking of word associations to *nobilis* in *Bon1lem*

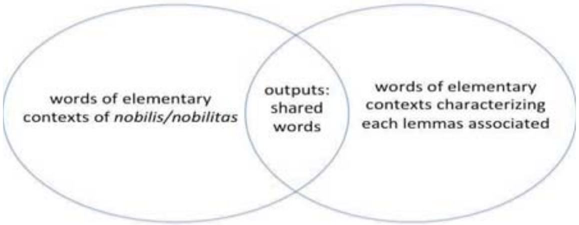
	LEMMA B	*EC(AB)*	*PROXIMITY*
1	*sum*	253	0,891
2	*nobilitas*	130	0,616
3	*Deus*	116	0,609
4	*habeo*	113	0,602
5	*ratio*	103	0,585
6	*video*	87	0,537
7	*possum*	86	0,504
8	*natura*	76	0,498
9	*corpus*	72	0,494
10	*anima*	72	0,488

Table 7: ranking of word associations to *nobilitas* in *Bon1lem*

	LEMMA B	*EC(AB)*	*PROXIMITY*
1	*sum*	164	0,726
2	*nobilis*	130	0,616
4	*Deus*	80	0,527
3	*ratio*	75	0,535
5	*habeo*	72	0,482
6	*possum*	58	0,427
7	*natura*	54	0,444
8	*video*	51	0,395
9	*objicio*	49	0,41
10	*creatura*	44	0,397

3.2. Comparison between Pairs of Key Words

For each of the first ten words most frequently associated to *nobilis* and *nobilitas* and included in the tables 4, 5, 6 and 7, we considered the terms characterizing their respective elementary contexts. In so doing, the proposed task was to verify which terms are the most shared with *nobilis* and *nobilitas*. The following draft illustrates our way of proceeding:


words of elementary contexts of *nobilis/nobilitas*
outputs: shared words
words of elementary contexts characterizing each lemmas associated

For example, the outputs of the analysis on the pairs of words *anima-nobilis* and *corpus-nobilis* are here transcribed[16]:

Table 8: anima-nobilis in Alb1lem

	LEMMA C	ASS C (AB)
1	sum	115
2	intelligentia	52
3	habeo	50
4	causa	48
5	corpus	42
6	natura	38
7	nobilitas	34
8	forma	33
9	operatio	28
10	virtus	27

Table 9: corpus-nobilis in Alb1lem

	LEMMA C	ASS C(AB)
1	sum	91
2	habeo	49
3	anima	42
4	natura	38
5	nobilitas	35
6	possum	29
7	causa	27
8	video	21
9	moveo	20
10	inferus	20

Table 10: anima-nobilis in Bon1lem

	LEMMA C	ASS C (AB)
1	sum	70
2	habeo	38
3	nobilitas	38
4	corpus	34
5	ratio	31
6	Deus	30
7	possum	25
8	natura	24
9	Christus	23
10	video	22

Table 11: corpus-nobilis in Bon1lem

	LEMMA C	ASS C(AB)
1	sum	70
2	habeo	35
3	anima	34
4	nobilitas	33
5	ratio	24
6	video	23
7	natura	22
8	forma	21
9	Deus	21
10	parvus	20

4. Results

In this work we have described the implementation of elementary association indexes on two distinct *corpora* of medieval texts (*Alb1lem* and *Bon1lem*), selected by assuming two peculiar KWIC, the adjective *nobilis* and the noun *nobilitas*. In this regard, some conclusive remarks can be formulated as follows

(1) Without having any particular acquaintance with Albert the Great's thought and his conception of *nobility*, one can assert that it plays presumably a meaningful function in his psychology and noetic. Among the ten most recurrently terms associated to *nobilis* and *nobilitas*, at least three of them, *anima*, *intelligentia* and *ratio* refer peculiarly to this epistemological field. Table 12 shows the total of EC shared by occurrences of our KWIC and the mentioned words:

Table 12: Noetical-Psychological Elementary contexts for Albert the Great's *nobility*

	Noetical-Psychological EC	*% (on the first ten word associations)*	*% (on the first ten word associations, without* sum and habeo)
Nobilis	177	16,76	25,14
Nobilitas	63	11,23	16,77

(2) In Albert the Great's inspected texts a remarkable number of elementary contexts of *nobilis* and *nobilitas* are associated to the words *anima* and *corpus*, which particularly characterize the anthropological debate.

Table 13: Anthropological debate EC for Albert the Great's *nobility*

	Anthropological debate EC	*% (on the first ten word associations)*	*% (on the first ten word associations, without* sum and habeo)
Nobilis	206	18,65	29,26
Nobilitas	63	12,83	19,09

(3) In the light of this last consideration, the second series of experiment was exemplified with outputs related to pair of terms *anima-nobilis* and *corpus-nobilis*. By matching the words shared by the two pairs of terms, the following table can be displayed:

Table 14: Words shared by the pair of words anima-nobilis/corpus-nobilis

anima-nobilis/ corpus-nobilis
Anima
Causa
Corpus
Forma
Habeo
Inferus
Intelligentia
Moveo
Natura
Nobilitas
Operatio
Sum
Video
Virtus

As the matter of facts, among the 14 words shared by the elementary contexts of *anima/nobilis* and *corpus/nobilis*, 8 of them (in italic) already emerged as word associated to *nobilis* in the *Table 4*. Now, excluding the already-quoted *intelligentia* and *ratio*, we have enough elements to suppose that also terms such as *causa, natura* are to be connected to an anthropological discussion.

(4) Contrasting with the habitual convinctions, *nobility* is not here principally connected to ethical or political fields. This seems to suggest that noetical and anthropological contexts could be a sort of starting point to interpret the uses of *nobility* proposed in other kind of medieval works. The analysis on Bonaventure's texts permits to formulate additional and comparative observations.

(4) Of course the most significance difference between Albert the Great's and Bonaventure's notion of *nobility* resides in the presence – among the word associated in the elementary contexts of *Bon1lem* – of terms such as

Deus and *creatura*. The fact that similar or analogous terms (*Deus, Christus*) emerge also in the analysis of the pairs of word (*anima-nobilis/corpus-nobilis*) corroborated our sensation. These peculiar associations can be connected to two kinds of reasons: (a) a different subject of the writings included in the two *corpora*; (b) a dissimilar conception of *nobility*.

(a) *Alb-corpus* includes several commentaries on Aristotle, while *Bon-corpus* is fundamentally composed by theological works (among the others the *Commentary on the Sentences*).

(b) According to the Bonaventure master the *nobility* is evidently related to a theological perspective that in Albert's writings seems to be absent, even if among the text included in *Alb1lem*, there are also theological writings, such as *De sacramentis*, *De resurrectione*, *Summa mirabili scientia Dei*.

5. Evaluation and conclusion

The starting point that we have assumed to test tools for an automatical text-processing of medieval works is *de facto* typical among the medievistic scholars: in order to circumscribe the significance and the function played by a concept within a group of texts (in our case, *nobilis* and *nobilitas*), all occurrences including the designated KWIC are taken into consideration. However, in the most cases, this initial textual approach is immediately neglected: after identifying a subset of the most meaningful occurrences with their possible sources, the analysis is usually addressed on a theoretical lecture of the data. In so doing, all lexical and semantic properties emerging from texts are not effectively capitalized. On the contrary, in our case, we have observed how an automatic or semi-automatic analysis on a group of texts can reveal interesting aspects of their subjects, without losing sight of the textual basis.

As the matter of facts, the elementary algoritms applied here represents only an example of the tools that could be exploited on a text-*corpus*. In any case, the advantages for a medievistic study are tangible:

- it is possible to establish the contexts of use of a particular notion and then to formulate well-founded suppositions regarding its significance;
- the subject and the contents of work or of a group of writings can be defined;
- differences and influences among two or more texts can be evaluated not only in the light of a statistical inventary of the shared occurrences, but also by comparing elementary semantic contexts.

Of course tricky aspects are not missing:

- the best results of this kind of analysis are obtained by considering terms not particularly frequent (for example, *nobilis* and *nobilitas*). By contrast, it would be very difficult to interpret the outputs of analysis on words more recurring such as *forma*, *esse*, *causa*, etc;
- at the same time, an examination of a very circumscribed number of occurrences offers presumably not so-faithful results. In fact, in these particular cases study, the notion of «frequency» should be considered in a relative and not in an absolute sense;
- for some words a possible transformation of meaning is to take into consideration;
- sometimes the occurrences are included in quotations or transcriptions and therefore they do not always represent the opinion of the author.

To conclude, the experiment reported here is evidently too restricted to give an exhaustive answer to the question raised at the beginning of our investigation (could computer analysis provide continuity between philology and philosophy in the medieval studies?) However, it offers indicative suggestions which are worth to be recapitulated. If a philological approach is considered more objective than a philosophical reflection, an automatic semantic analysis of the texts gives furthermore an undertaking of this prerequisite but also highlights a series of new information emerging from the lexical units, which display something about the contents of a *corpus* of writings. Co-occurrences, Word associations, Indexes of proximity are only some of the elementary outputs on which any scholar of medieval philosophy can base additional impartial considerations on the lexical properties of a text, before expressing his theoretical convinction. Starting from these premises, in the future, it will be possible to design automatic tools for the attribution of anonymous medieval texts in the light of the lexical similarities among the elementary contexts. But, by adopting the same approach, we cannot completely exclude that computer contributes to achieve also more elaborate information: for example, discriminating the way of raising a problem or of formulating it among two or more *corpora* of texts. These suggestions sound futuristic, but also in many other circumstances we are providing room for computers. Perhaps they will become the new narrators of the history of medieval philosophy.

Note

[1] This study is a partial development of the research project *Testi medievali e mappe ontologiche digitali. Il concetto di* nobilitas *come* speculum *per una web-analysis delle teorie dell'intelletto del XIII secolo*, submitted under a post-doc fellowship in 2011 (Provincia Autonoma di Trento). I would like to thank Prof. Alfio Ferrara (University of Milan) for his precious corrections and comments.

[2] Among the numerous examples of databases, my recent *DB Anima* (http://nobilitas.lett.unitn.it/ strumenti.html) is worth here to be mentioned.

[3] In this perspective, several international research projects have been undertaken. See, especially, *Cost Action IS1005* (http://www.medioevoeuropeo.eu/).

[4] Albertus Magnus, *Opera omnia, editio Coloniensis*, Aschendorff, Münster 1951-2004.

[5] Bonaventura, *Opera omnia*, Ad Claras Aquas (Quaracchi), prope Florentiam : ex Typographia Collegii S. Bonaventurae, 1882-1902

[6] http://demo.albertus-magnus-online.de/

[7] http://clt.brepolis.net/llta/Default.aspx

[8] Distinction between *argumenta ad* and *contra* is not here taken into consideration.

[9] Stopwords are not included.

[10] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[11] http://tlab.it/en/presentation.php

[12] T-LAB considers an elementary context to be every sequence of words interrupted by full stop and carriage return, whose dimensions are inferior to 400 characters. in the case where, within the maximum length, a full stop is not present, it searches for other punctuation marks in the following order (? ! ; : ,). If none are found, it performs segmentation on the basis of a statistical criterion, but without cutting the lexical units.

[13] Co-occurrences, then, are quantities which result from a computation of how many times two or more lexical units are present together in the same elementary contexts (EC).

[14] In T-LAB the association indexes (or similarity coefficients) are used to analyse the co-occurrences of the lexical units (LU) inside the elementary contexts (EC), that is to analyse binary data of the presence/absence type. In this specific circumstance, we adopted cosine formula: $\frac{a}{\sqrt{(a+b)}\ X\ \sqrt{(a+c)}}$., *Tools for Text Analysis. User's Manual*, http://tlab.it/it/download.php, p. 200

[15] Ranking is expressed in the first column. LEMMA B includes the words associated to *nobilis* (tables 4 and 6) and to *nobilitas* (tables 5 and 7). EC (AB) includes the number of elementary contexts shared by *nobilis* (tables 4 and 6) or *nobilitas* (tables 5 and 7) with LEMMA B. PROXIMITY indicates an equivalence index obtained by dividing their squared co-occurrences by the product of their occurrences. Data were classified as decreasing EC (AB).

16 Ranking is expressed in the first column. LEMMA includes the words most frequently shared between the elementary contexts of *anima* and *nobilis* (tables 8 and 10) and of *corpus* and *nobilis* (tables 9 and 11). ASS (AB) includes the number of elementary contexts shared. Data were classified as decreasing ASS (AB)*.* Cosine is the adopted formula.