UPO

UNIVERSITÀ DEL PIEMONTE ORIENTALE

# Past and present in the use of computer for the analysis of texts

Maurizio Lana

Università del Piemonte Orientale

# science and its construction science and its narration

- in day-to-day research — what Bruno Latour calls *science in the making* — far from simply discovering facts, scientists seem to be "in the business of being convinced and convincing others." During the process of arguing over uncertain data, scientists foregrounded the reality that they were speaking *for* the facts; and yet, as soon as their propositions were turned into indisputable statements and peer-reviewed papers — what Latour calls *ready-made science* — they claimed that such facts had spoken for themselves. That is once the scientific community accepted something as true were the all-too-human processes ehind it effectively erased or, as Latour put it, black-boxed
  from: https://www.nytimes.com/2018/10/25/magazine/bruno-latour-post-truth-philosopher-science.html#click=https://t.co/G2cT4j8Pb0
- hence narration is a proper way to introduce to / to enter into science

# after Latour’s, my words

- science is a social construction, a network of person and ideas, not only results
- results are emergences which hide the network, the underlying part of the iceberg
- if you don’t know the existence of the underlying part ... Titanic! you crash against the iceberg

= you don't understand the complexity of the field and possibly misuse its tools

# two networks and two narrations

- at least two main networks and related narratives
- a euro-american narrative: Busa – IBM and what follows
- a US-based narrative: CATSS* and Pennsylvania University

*CATSS=Computer Assisted Tools for Septuagint Studies

# religious studies

- in both cases initial focus on religious texts: for Busa theological texts: the Summa Theologica of Thomas Aquinas
- for CATSS the Bible: the Septuagint Bible, Greek translation of Hebrew bible

# religious studies, not casually...

- ...I think. the Christian and the Hebrew religions are so-called book religions, based on the book of the Bible
- the Bible is the first hypertext in the history of our culture: more recent books cite *by words* what is said in previous books.
- this produced a study of the Bible based on the study of its words

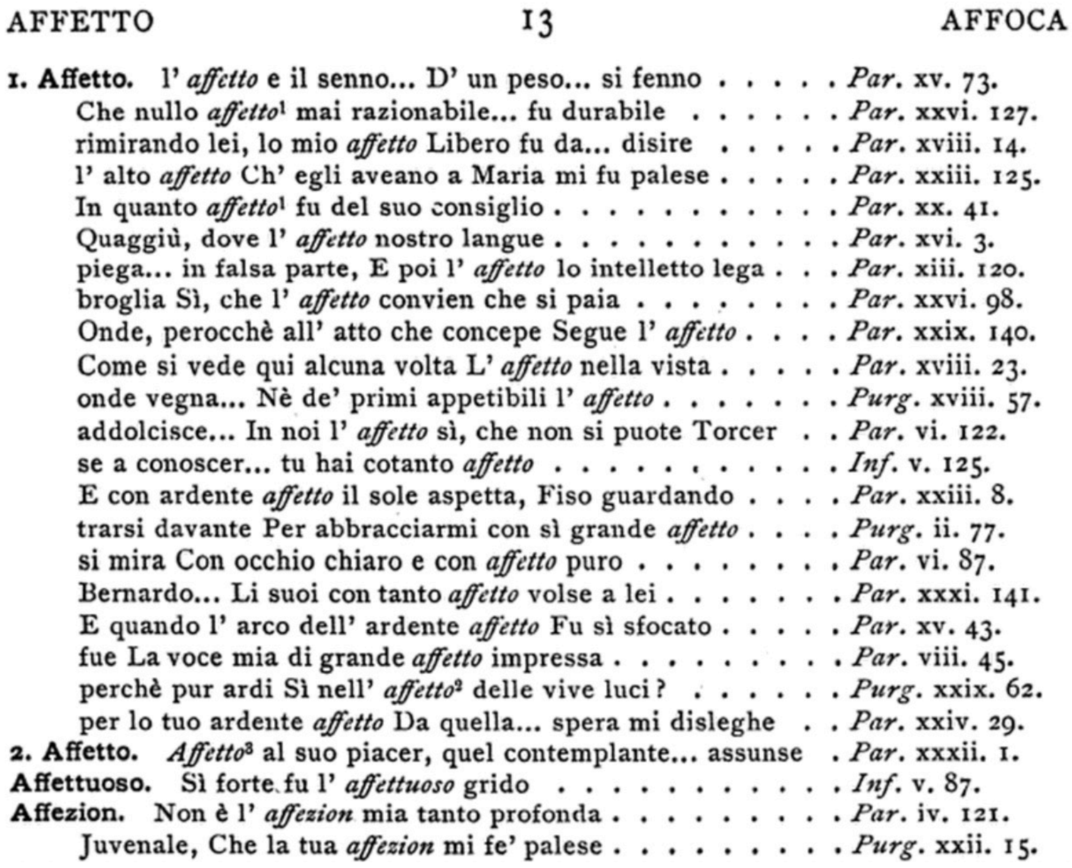
  which are God’s words, by the way, hence deserve a maximum level of scientific devotion

# where/when the concordances were born

- the *concordance* was invented in Paris, abbaye de St. Jacques, under the guidance of Hugues de Saint-Cher, in 1230 on the basis of the Latin Vulgate

source: https://big.hypotheses.org/1928

# a *modern* concordance to Dante, Divina Commedia



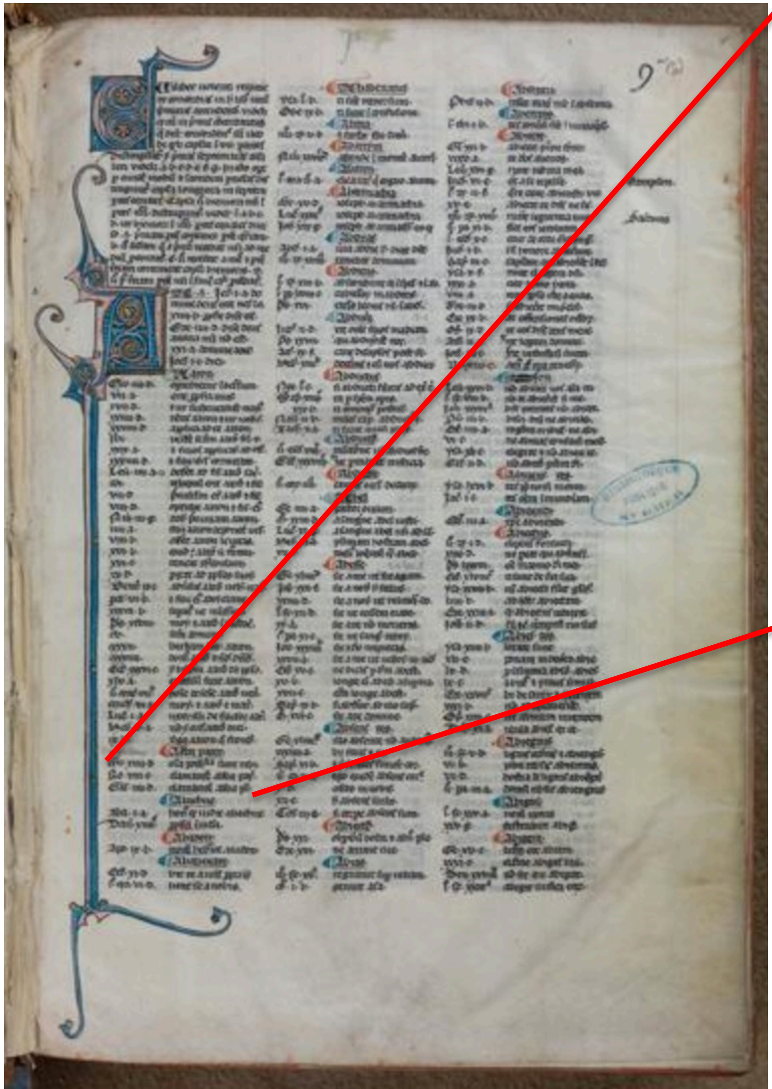
AFFETTO 13 AFFOCA

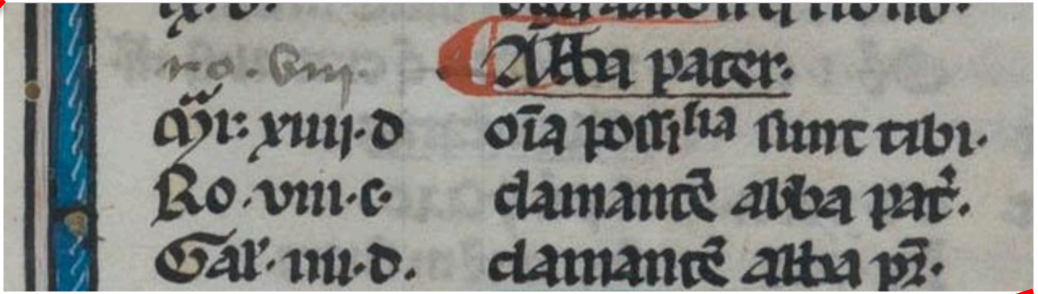
| | | |
|---|---|---|
| **1. Affetto.** | l' *affetto* e il senno... D' un peso... si fenno . . . . . | *Par.* xv. 73. |
| | Che nullo *affetto*[1] mai razionabile... fu durabile . . . . . . | *Par.* xxvi. 127. |
| | rimirando lei, lo mio *affetto* Libero fu da... disire . . . . . | *Par.* xviii. 14. |
| | l' alto *affetto* Ch' egli aveano a Maria mi fu palese . . . . . | *Par.* xxiii. 125. |
| | In quanto *affetto*[1] fu del suo consiglio . . . . . . . . . . | *Par.* xx. 41. |
| | Quaggiù, dove l' *affetto* nostro langue . . . . . . . . . . . | *Par.* xvi. 3. |
| | piega... in falsa parte, E poi l' *affetto* lo intelletto lega . . . | *Par.* xiii. 120. |
| | broglia Sì, che l' *affetto* convien che si paia . . . . . . . . | *Par.* xxvi. 98. |
| | Onde, perocchè all' atto che concepe Segue l' *affetto* . . . . | *Par.* xxix. 140. |
| | Come si vede qui alcuna volta L' *affetto* nella vista . . . . . | *Par.* xviii. 23. |
| | onde vegna... Nè de' primi appetibili l' *affetto* . . . . . . . | *Purg.* xviii. 57. |
| | addolcisce... In noi l' *affetto* sì, che non si puote Torcer . . | *Par.* vi. 122. |
| | se a conoscer... tu hai cotanto *affetto* . . . . . . . . . . . | *Inf.* v. 125. |
| | E con ardente *affetto* il sole aspetta, Fiso guardando . . . . | *Par.* xxiii. 8. |
| | trarsi davante Per abbracciarmi con sì grande *affetto* . . . . | *Purg.* ii. 77. |
| | si mira Con occhio chiaro e con *affetto* puro . . . . . . . . | *Par.* vi. 87. |
| | Bernardo... Li suoi con tanto *affetto* volse a lei . . . . . . . | *Par.* xxxi. 141. |
| | E quando l' arco dell' ardente *affetto* Fu sì sfocato . . . . . | *Par.* xv. 43. |
| | fue La voce mia di grande *affetto* impressa . . . . . . . . . | *Par.* viii. 45. |
| | perchè pur ardi Sì nell' *affetto*[2] delle vive luci ? . . . . . . | *Purg.* xxix. 62. |
| | per lo tuo ardente *affetto* Da quella... spera mi disleghe . . | *Par.* xxiv. 29. |
| **2. Affetto.** | *Affetto*[3] al suo piacer, quel contemplante... assunse . | *Par.* xxxii. 1. |
| **Affettuoso.** | Sì forte fu l' *affettuoso* grido . . . . . . . . . . . | *Inf.* v. 87. |
| **Affezion.** | Non è l' *affezion* mia tanto profonda . . . . . . . . . . | *Par.* iv. 121. |
| | Juvenale, Che la tua *affezion* mi fe' palese . . . . . . . . . | *Purg.* xxii. 15. |

# the concordance of Hugues de St. Cher

- for every word of the Bible, the passages containing it where indicated by book and chapter but not by verses (first introduced in 1545 by Robert Estienne); Hugo divided each chapter into seven almost equal parts, indicated by the letters of the alphabet, a, b, c, etc.

# the concordance of Hugues de St. Cher





| | Abba pater |
|---|---|
| [marcus]xiiii.d | omnia possibilia sunt tibi |
| Ro.viii.c | clamantes abba pater |
| Gal.iiii.d | clamantes abba pater |

# the key concept of concordance

- explain the text with the text itself

  that is: collect the passages which concord in the use of a same word

- the conception of the text as a universe with its internal rules which the reader must understand in order to grasp the meaning of the textual universe

# explain the text with the text itself

- and consequently:
  the meaning of a word can be explained with the passages where it recurs
- read the passages where a same word recurs, as connected to each other

- back to the two lines / two networks / two narratives:
  Busa
  CATSS

# Busa endeavour

- to build a *new* concordance of the Summa Theologica of Thomas Aquinas showing the *in* preposition not recorded in the existing printed concordances (aim: studying the concept of *esse in*...)
- after manual collection of 10thousand occurrences the task was too big to manage that way son in 1949 he went to Thomas Watson, the chair of IBM, a company which produced computers, and with their help he managed the task

  NB: those times computers computed, that is, they did *calculations with numbers!*
- **key concept**: we use computers to do what we can conceive but which requires too much time

# digital and print

- final printed product:
  Index Thomisticus. Sancti Thomae Aquinatis operum omnium indices et concordantiae (Stuttgart, Frommann Holzboog, 1974-1980)

  56 volumes, 62thousand pages, every single word used in every work of Th. Aq. is indexed
- all of this work was subsequently put onto 1 CD

  what means that the whole corpus of Thomas Aquinas has very small dimensions if we think of it in terms of "raw data"

# big data in the study of texts?

- 1 CD = 640.000.000 bytes / 640 millions bytes / 640MB
- 1 single experiment at CERN LHC (Large Hadron Collider) produces
  600.000.000.000.000 of bytes per second
  600 millions of MBytes per second
  nearly 1 million of CDs per second
  source https://home.cern/about/computing/processing-what-record
- **key concept**: no corpus of textual data can be assimilated to what today is called big data

# Busa accomplishment

- Busa made clear that computers could be used to compute textual data
- compute: internally, a CPU central processing unit, is a 'place' where only calculations happen

  what in turn shows the terrific power of the binary system (arithmetic, logic, etc.) and the connected technology (the computers): it represents words and works on them using only two digits, 0 and 1

# textual data

- **key concept**: the *text* is broken down into elements which are *data*

  the text is dismantled into its constituents, its structures tend to be dissolved

- elements: their granularity level is decided by the scholar

  chapters, phrases, clauses of every type, metaphors, words, roots / endings and other phonic/phonetic components, …

# second narrative: CATSS

- **CATSS** Computer Assisted Tools for Septuagint Studies was based in Pennsylvania State University

source for details https://www.sas.upenn.edu/~jtreat/rs/rscpuhx.html

- ENIAC, the second complete Turing electronic computer was created at Pennsylvania State University in 1946

# CATSS

- from 1968 IOSCS - International Organization for Septuagint and Cognate Studies - wanted to create a lexicon for the Greek Jewish Scriptures
- in 1972 Robert Kraft of CATSS (Pennsylvania University) suggested to use computer technology to store and analyze the data
- feasibility study by Jack Abercrombie, who had computing experience and was writing his thesis in archaeology at Pennsylvania University

he visited computing centers: by Emanuel Tov in Jerusalem, F. Poswick at Maredsous (Belgium), the Septuagint Project in Göttingen, and Susan Hockey at the Oxford Text Archive.

# the community of computers in the humanities

- great impact  of the project of Septuagint Lexicon on the community of "computing in the humanities" which was building around the Humanist mailing list

  the work of Busa and his group didn't really entered this realm.

- the project connected with/intercepted the growing interest for the use of computers expressed by classical scholars

  which produced / was created by the TLG cdrom (collection of all the classical and byzantine Greek works) and the PHI cdrom (collection of all the archaic an classical latin works) and by the Ibycus workstation by David Packard

# the community of computers in the humanities

- the PHI and TLG cdrom greatly fostered the *curiosity* for the text analysis of ancient works
  - they gave birth to a bunch of concordance software: TLG workstation, PHI workstation, OFFLOAD, Lector, SNS-Greek and Latin, Accordance, Hellespont, all now gone.
  - and Diogenes, both for PC and Mac, which is alive and kicking, by Peter Heslin

# types of concordances

we can conceive 4 type of concordances:

- verbal - index of occurences without context
- contextual - KWIC (Key Word In Context) most frequent
- lemmatized - word-forms are subsumed into their base-form (the lemma)
- conceptual - organized by idea or sense
  - build a group of words connected by whichever reason (usually meaning): house, home, palace, barn, hut, tent, …; give the group a name, say "home”, and then search for the group
  - this could be done through a thesaurus where words are placed in structures of synonymy/hyponymy/ hyperonymy

# from mainframes to PCs

- the text analysis which initially was made with mainframes – big departmental computers – crossed the path of the personal computers

# from mainframes to PCs

- programs for doing indexes and concordances with personal computer were produced
- what meant that you were able to do by your own a work which traditionally required lots of time by lots of people

to do this you need only: one or more texts in a *format readable by the computer* and a program to read the file, find the words, extract them and present them on the output device -screen or printer

# CoCoA, OCP, MicroOcp

- OCP / MicroOCP owed to CoCoA Count and COncordances on Atlas
  http://www.chilton-computing.org.uk/acl/applications/cocoa/p001.htm
- the info about the structure was inserted in the text in this form:

  <W SHAKESPEARE> <T HAMLET> <A 1> <S 1> plus L for line, directly managed by the program

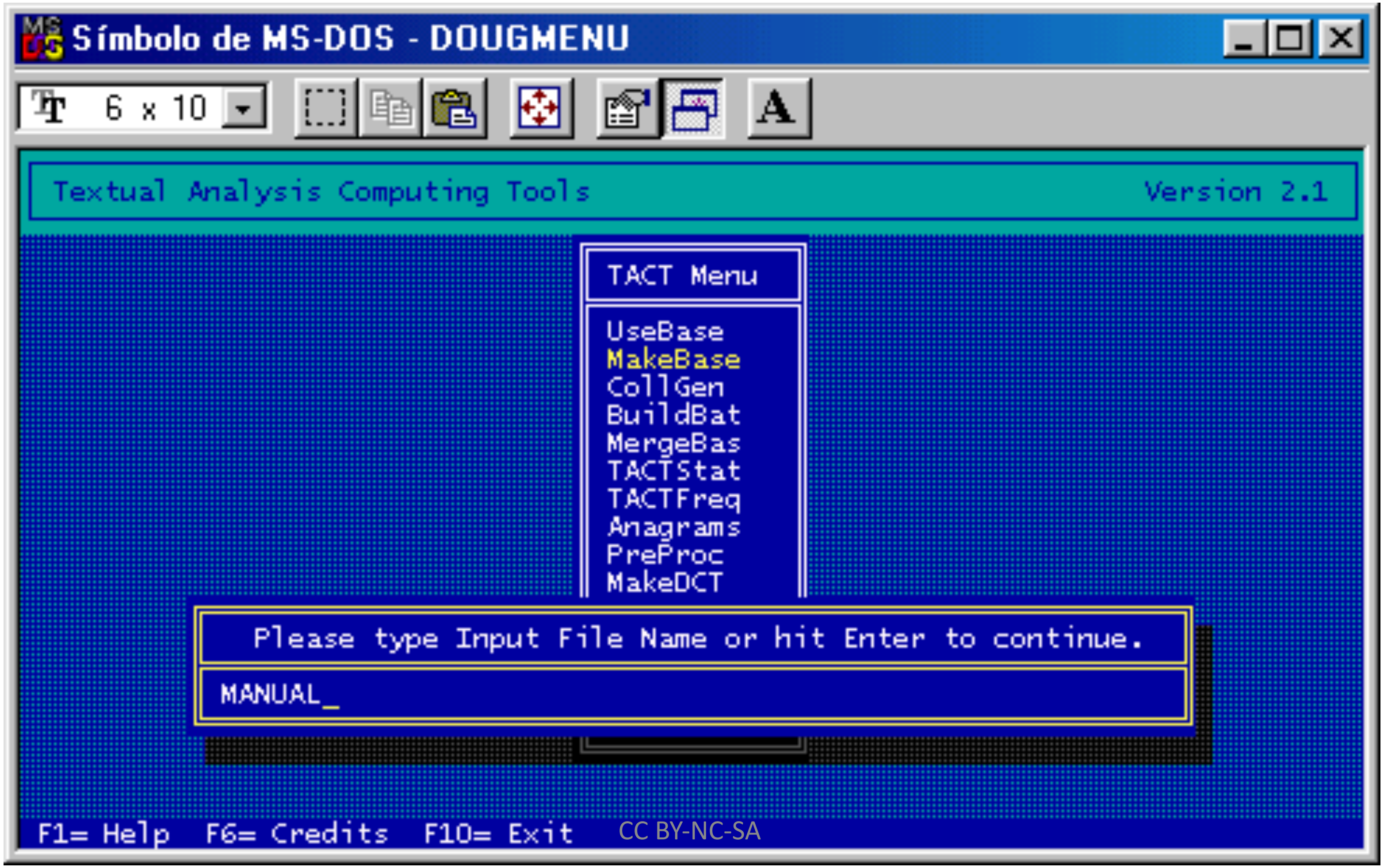
# WordCruncher

- then came Wordcruncher
- still alive today, marginal
- was born in the Brigham Young University in Utah and was promoted with sample texts by Shakespeare and from the Bible.

# TACT

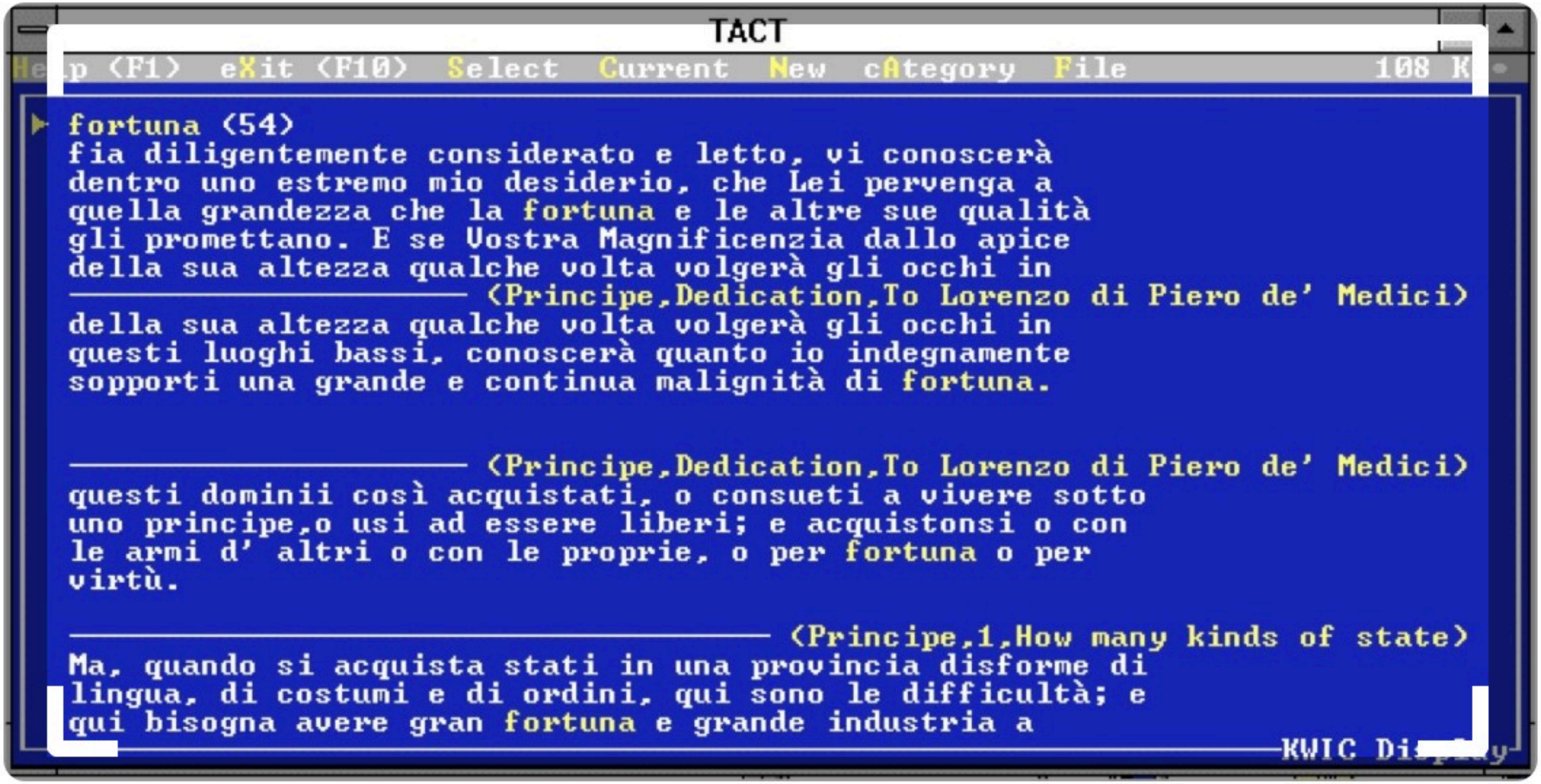
- TACT - Text Analysis Computing Tools from University of Toronto was the first software decently accompanied by a manual which was also a book about the study of texts with computing tools
- see http://projects.chass.utoronto.ca/tact/TACT/tact0.html

# TACT



```
Textual Analysis Computing Tools                                    Version 2.1

                                 TACT Menu

                                 UseBase
                                 MakeBase
                                 CollGen
                                 BuildBat
                                 MergeBas
                                 TACTStat
                                 TACTFreq
                                 Anagrams
                                 PreProc
                                 MakeDCT

      Please type Input File Name or hit Enter to continue.
    MANUAL_

F1= Help   F6= Credits   F10= Exit
```

# TACT

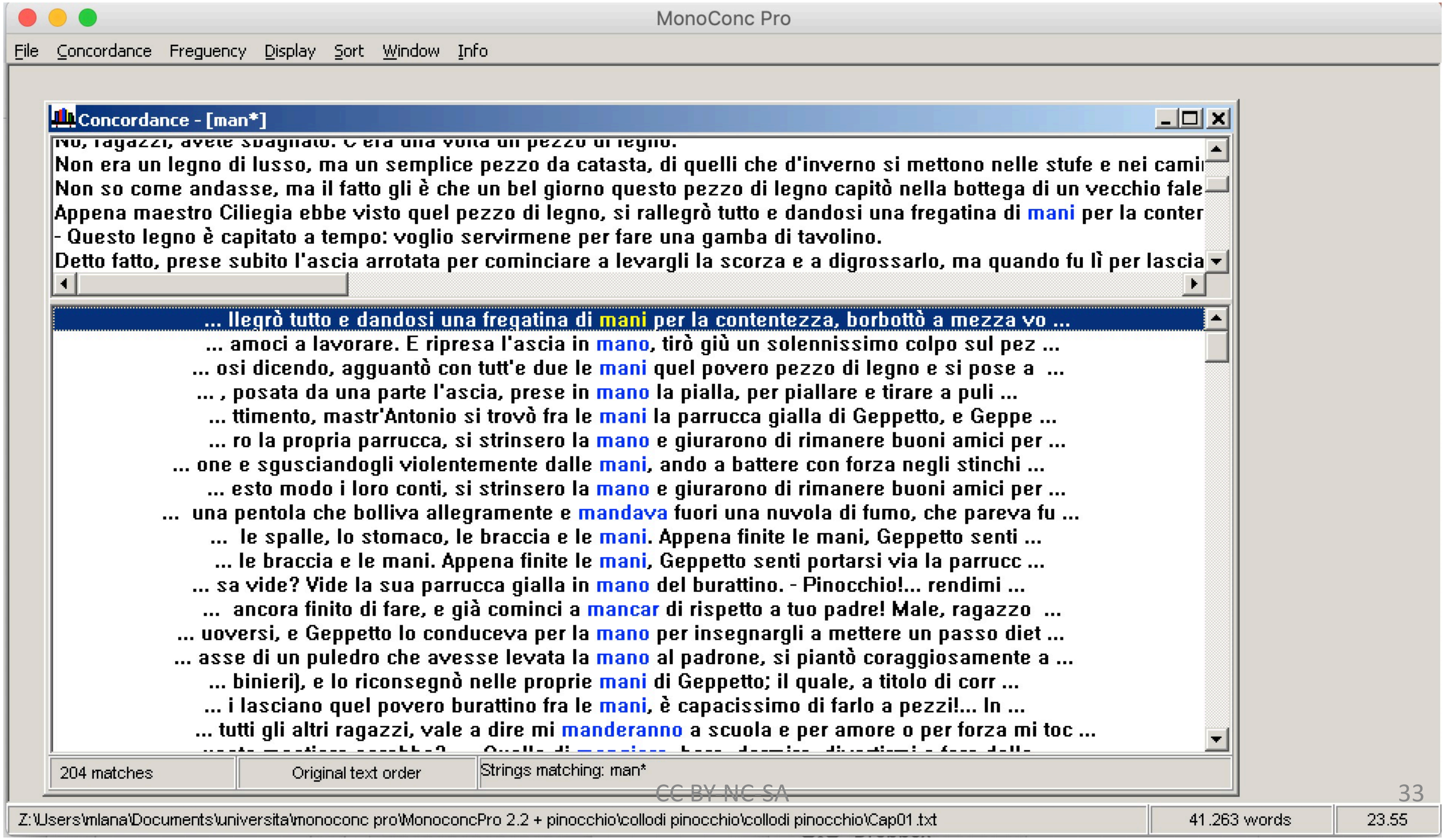


```
fortuna (54)
fia diligentemente considerato e letto, vi conoscerà
dentro uno estremo mio desiderio, che Lei pervenga a
quella grandezza che la fortuna e le altre sue qualità
gli promettano. E se Vostra Magnificenzia dallo apice
della sua altezza qualche volta volgerà gli occhi in
--------------------- (Principe,Dedication,To Lorenzo di Piero de' Medici)
della sua altezza qualche volta volgerà gli occhi in
questi luoghi bassi, conoscerà quanto io indegnamente
sopporti una grande e continua malignità di fortuna.


--------------------- (Principe,Dedication,To Lorenzo di Piero de' Medici)
questi dominii così acquistati, o consueti a vivere sotto
uno principe,o usi ad essere liberi; e acquistonsi o con
le armi d' altri o con le proprie, o per fortuna o per
virtù.

------------------------------------- (Principe,1,How many kinds of state)
Ma, quando si acquista stati in una provincia disforme di
lingua, di costumi e di ordini, qui sono le difficultà; e
qui bisogna avere gran fortuna e grande industria a
```

# MonoConc

- Monoconc doesn't require any pre-tagging to produce meaningful output
- apparently focused onto linguistic studies (corpus linguistics)
- http://www.athel.com/mono.html

# MonoConc



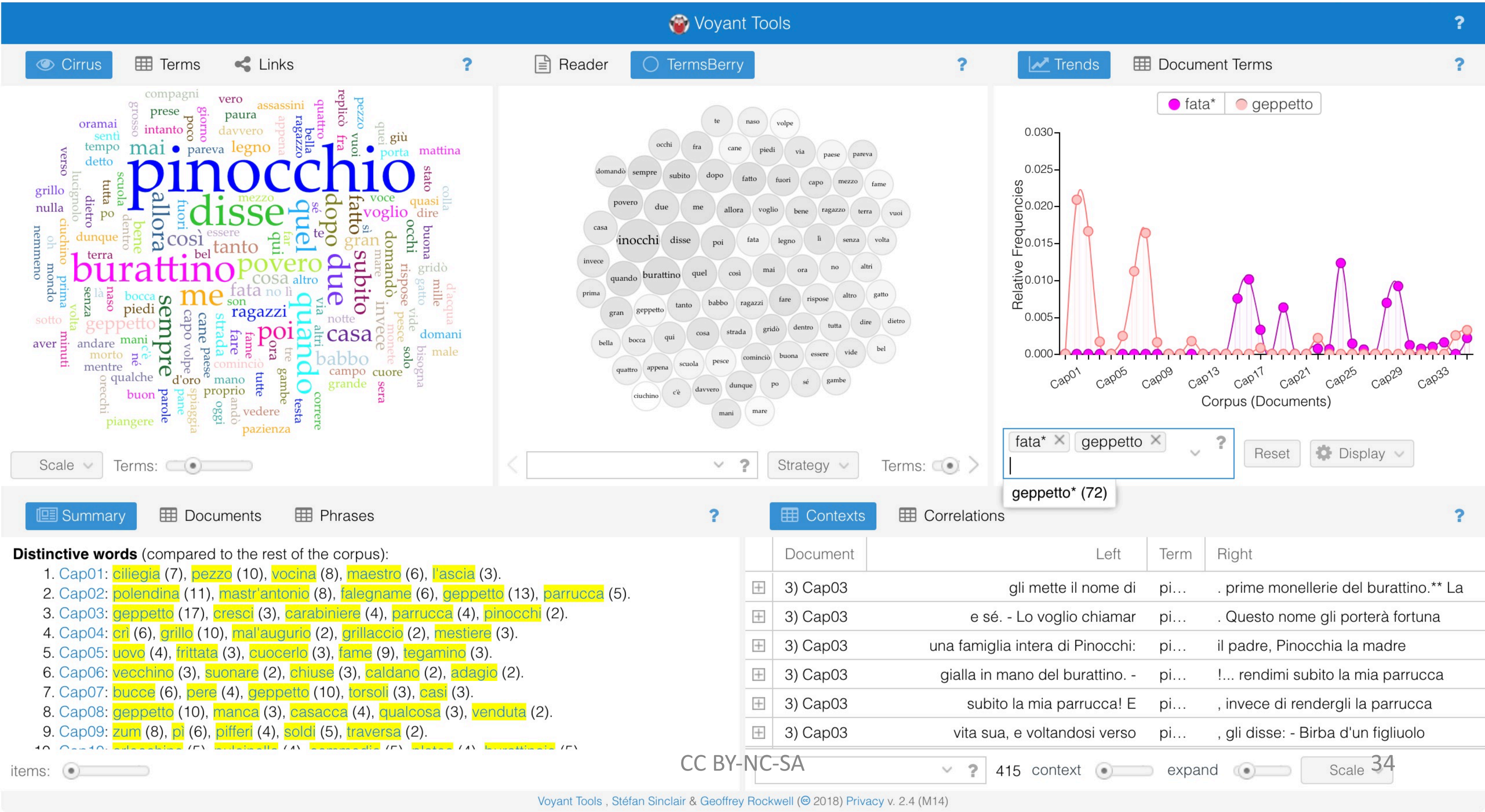
No, ragazzi, avete sbagliato. C'era una volta un pezzo di legno.
Non era un legno di lusso, ma un semplice pezzo da catasta, di quelli che d'inverno si mettono nelle stufe e nei cami
Non so come andasse, ma il fatto gli è che un bel giorno questo pezzo di legno capitò nella bottega di un vecchio fale
Appena maestro Ciliegia ebbe visto quel pezzo di legno, si rallegrò tutto e dandosi una fregatina di mani per la conter
- Questo legno è capitato a tempo: voglio servirmene per fare una gamba di tavolino.
Detto fatto, prese subito l'ascia arrotata per cominciare a levargli la scorza e a digrossarlo, ma quando fu lì per lascia

```
... llegrò tutto e dandosi una fregatina di mani per la contentezza, borbottò a mezza vo ...
... amoci a lavorare. E ripresa l'ascia in mano, tirò giù un solennissimo colpo sul pez ...
... osi dicendo, agguantò con tutt'e due le mani quel povero pezzo di legno e si pose a ...
... , posata da una parte l'ascia, prese in mano la pialla, per piallare e tirare a puli ...
... ttimento, mastr'Antonio si trovò fra le mani la parrucca gialla di Geppetto, e Geppe ...
... ro la propria parrucca, si strinsero la mano e giurarono di rimanere buoni amici per ...
... one e sgusciandogli violentemente dalle mani, ando a battere con forza negli stinchi ...
... esto modo i loro conti, si strinsero la mano e giurarono di rimanere buoni amici per ...
... una pentola che bolliva allegramente e mandava fuori una nuvola di fumo, che pareva fu ...
... le spalle, lo stomaco, le braccia e le mani. Appena finite le mani, Geppetto senti ...
... le braccia e le mani. Appena finite le mani, Geppetto senti portarsi via la parrucc ...
... sa vide? Vide la sua parrucca gialla in mano del burattino. - Pinocchio!... rendimi ...
... ancora finito di fare, e già cominci a mancar di rispetto a tuo padre! Male, ragazzo ...
... uoversi, e Geppetto lo conduceva per la mano per insegnargli a mettere un passo diet ...
... asse di un puledro che avesse levata la mano al padrone, si piantò coraggiosamente a ...
... binieri), e lo riconsegnò nelle proprie mani di Geppetto; il quale, a titolo di corr ...
... i lasciano quel povero burattino fra le mani, è capacissimo di farlo a pezzi!... In ...
... tutti gli altri ragazzi, vale a dire mi manderanno a scuola e per amore o per forza mi toc ...
... uesto mestiere sarebbe? - Quello di mangiare, bere, dormire, divertirmi e fare dalla ...
```

204 matches | Original text order | Strings matching: man*

Z:\Users\mlana\Documents\universita\monoconc pro\MonoconcPro 2.2 + pinocchio\collodi pinocchio\collodi pinocchio\Cap01.txt | 41.263 words | 23.55

# Voyant Tools



**Distinctive words** (compared to the rest of the corpus):

1. Cap01: ciliegia (7), pezzo (10), vocina (8), maestro (6), l'ascia (3).
2. Cap02: polendina (11), mastr'antonio (8), falegname (6), geppetto (13), parrucca (5).
3. Cap03: geppetto (17), cresci (3), carabiniere (4), parrucca (4), pinocchi (2).
4. Cap04: crì (6), grillo (10), mal'augurio (2), grillaccio (2), mestiere (3).
5. Cap05: uovo (4), frittata (3), cuocerlo (3), fame (9), tegamino (3).
6. Cap06: vecchino (3), suonare (2), chiuse (3), caldano (2), adagio (2).
7. Cap07: bucce (6), pere (4), geppetto (10), torsoli (3), casi (3).
8. Cap08: geppetto (10), manca (3), casacca (4), qualcosa (3), venduta (2).
9. Cap09: zum (8), pì (6), pifferi (4), soldi (5), traversa (2).
10. Cap10: arlecchino (5), pulcinella (4), commedia (5), platea (4), burattinaio (5)

| | Document | Left | Term | Right |
|---|---|---|---|---|
| ⊞ | 3) Cap03 | gli mette il nome di | pi… | . prime monellerie del burattino.** La |
| ⊞ | 3) Cap03 | e sé. - Lo voglio chiamar | pi… | . Questo nome gli porterà fortuna |
| ⊞ | 3) Cap03 | una famiglia intera di Pinocchi: | pi… | il padre, Pinocchia la madre |
| ⊞ | 3) Cap03 | gialla in mano del burattino. - | pi… | !... rendimi subito la mia parrucca |
| ⊞ | 3) Cap03 | subito la mia parrucca! E | pi… | , invece di rendergli la parrucca |
| ⊞ | 3) Cap03 | vita sua, e voltandosi verso | pi… | , gli disse: - Birba d'un figliuolo |

415

Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2018) Privacy v. 2.4 (M14)

# Voyant Tools

- they run online
- they are the best concordance environment we have today
- a rich manual which is also a book to read about textual analyses with digital tools: *Hermeneutica: computer-assisted interpretation in the humanities*, MIT Press, by Rockwell and Sinclair
- like TACT before: both tools come from the same canadian scholars environment

# the past

- this was the past even in very advanced form, as it is the case for the Voyant Tools: we continue the same activity invented in 1250 by Hugues de St. Cher

# the present
# with a foot in the future

- we pour (our) knowledge into the texts we want to study
- to do this we use annotation/markup/tagging
- the main (not unique) way to do this today is TEI annotation

possibly powered by true ontology-based annotation

# what is TEI

- TEI – Text Encoding Initiative dates back to 1987 but its meaning and usefulness clearly appeared only in the last 10-15 years
- aim: define how to formally describe every aspect of a text – structure and content
- the markup manages not only formal aspects but also content

“TEI guidelines” manual has more than 1900 pages

# content annotation in TEI

- network of music concepts in the "de musica" of Boetius

```
<interp xml:id=”chord”>chorda</interp>

<interpGrp type=”tetrarch”>
    <interp xml:id=”tetrach”>tetrachordum</interp>
    <interp xml:id=”diaton”>diatonum</interp>
    <interp xml:id=”chrom”>chroma</interp>
    <interp xml:id=”enarm”>enarmonium</interp>
</interpGrp>

<interpGrp type=”interv”>
    <interp xml:id=”interv”>intervallum</interp>
    <interp xml:id=”ton”>tonus</interp>
    <interp xml:id=”semiton”>semitonium</interp>
</interpGrp>

<interp xml:id=”son”>sonum</interp>

<interp xml:id=”conson”>consonantia</interp>
```

source: BA thesis of F. Michelone

# content annotation in TEI

- annotation of the text of "de musica" of Boetius

```
<p><seg ana="#interv">Nam si vox voce duplo sit acuta vel
gravis, diapason consonantia fiet, si vox voce sesqualtera
proportione sit vel sesquitertia vel sesquioctava acutior graviorque,
diapente vel diatessaron vel tonum consonantiam reddet.
Diatessaron igitur ac diapente unam diapason concinentiam
iungunt. </seg> <seg ana="#ton">Atque ut i d facillime
conprobetur, sit sesquioctava proportio VIII et VIIII Horum nullus
naturaliter medius numerus incidet. </seg> <seg
ana="#semiton"> Sed utraque semitonia nuncupantur, non
quod omnino semitonia ex aequo sint media, sed quod semum dici
solet, quod ad integritatem usque non pervenit. Sed inter haec
unum maius semito-nium nuncupatur, aliud minus.
</seg></p></div>
```

source: BA thesis of F. Michelone

# ontology

- in the previous slides we have an *elementary* computer ontology: a formalized description of a knowledge domain allowing both humans and machines to use it

# a more complex ontology

- GO! - Geographical Ontology describes the geographical knowledge contained in classical latin texts and is made of 4 basic ontologies:

  it is written in OWL - Web Ontology Language
- GO-TOP: main geographical concepts
- GO-PHY: description of physical geography
- GO-HUM: description of anthropic geography
- GO-FAR (For Ancient Resources): description of entities specific of classical world (see for example Ades, vallum, agora, etc.)

# GO! in action

pax ita conuenerat ut Etruscis Latinisque fluuius<geogName ref='http://www.geolat.it/geoData/Tiber'xml:id='tiber01> Albula</geogName>, quem nunc<geogName ref='http://www.geolat.it/geoData/Tiber’ xml:id='tiber02’>Tiberim </geogName> uocant, finis esset

- <geogName> identifies a noun or a phrase referring to a physical place
- xml:id univocally identifies the encoded element
- ref includes the URI which identifies the geographical entity referred by the text passage in GSD

# final words

- much can be done with computers to study texts
- but everything start with the simple act of reading a text through a concordance

thank you!