

From manuscript to syntactic tree: the long journey of mathematical Latin

Margherita Fantoli (margherita.fantoli@kuleuven.be)

Miryam de Lhoneux (ml@di.ku.dk)

Beatrice Sisana (beatrice.sisana@uniroma3.it)

Structure of the presentation

- Introduction
 - General goal of the project
- Part I: The corpus
 - The work of Archimedes
 - Jacopo's translation
 - MauroTeX
- Part II: Linguistic annotation
 - POS tagging
 - Syntactic tree
- Part III: Training a parser
 - Results
 - Error analysis

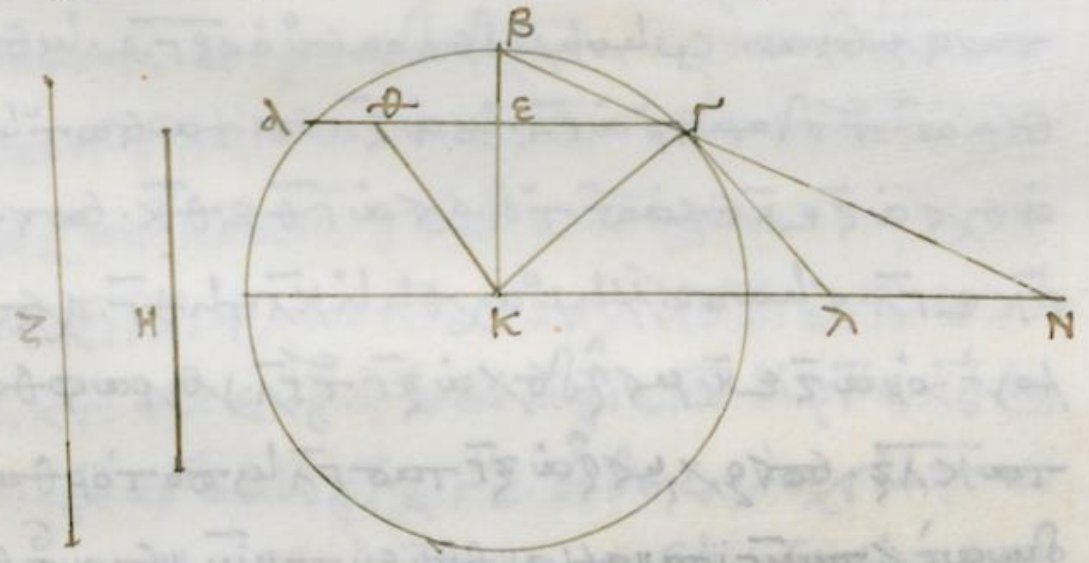
Goal of the project

- Create (in a semi-automated way) a treebank of the (Neo)Latin translation of Archimedes' texts
- Why?
 - **Historical interest:** key-translation for the dissemination of Archimedes' results in the West
 - **Linguistic interest:** a very specific variety of language (mathematical “natural” language) not extensively studied yet
 - **Corpus interest:** lack of resources and guidelines for the annotation of non-classical and non-literary Latin
 - **NLP interest:** do treebank embeddings work well for genre-specific parsing?

Archimedes' corpus

- The works of Archimedes were originally written in Ancient Greek and transmitted by a few manuscripts that reached Western Europe in the middle Ages (or were discovered in the XX century)
- Greek mathematics was a 'literary' genre on its own: specific vocabulary and stylistic conventions
- Texts were written in full natural language: no symbolic notation as we know it today was used, but diagrams are present in the extant copies
- At the moment, 11 of Archimedes' works have reached us

γάρ. εἰς τὸ αὐτὸ γὰρ ποτιταθῆκε. γὰρ κγ. ποτιταθῆκε. ἴσους γὰρ
 ὄντας αὐτὸ ποτιταθῆκε. ἴσους γὰρ κγ ποτιταθῆκε. ὅρ δὲ μὲν ὄντας αὐτὸ ποτι
 ταθῆκε. ἴσους γὰρ αὐτὸ κγ εἰς μὲν γὰρ ποτιταθῆκε βγ. ἴσους γὰρ αὐτὸ βγ.
 μεταξὺ τῶν περιφραγῶν ἡ τῶν ἀφραγῶν διατῆρε. διατῆρε γὰρ τὸ γμ.
 ἡ ἀφραγῶν ἡ τῶν
 περιφραγῶν
 γλ. ὄντας αὐτὸ κβγ εἰς βγ
 τῶν δὲ μὲν ὄντας αὐτὸ
 ποτιταθῆκε. ἴσους γὰρ αὐτὸ
 βγ. τῶν δὲ μὲν ὄντας αὐτὸ
 ὄντας αὐτὸ ποτιταθῆκε:



Laurenziano Plut. 28.04, f . 74v (XV century)

Example of literal translation of an Archimedes' sentence: So, there is a certain circle, ABΓ, its center K, and a line in the circle, smaller than the diameter, ΓΑ, and the ratio, which Z has to H, smaller than the <ratio> which ΓΘ has to ΚΘ, the perpendicular to it <=to ΓΑ> drawn; drawn from center the line ΚΝ, parallel to line ΓΑ et line ΓΛ perpendicular to ΚΓ

The translation of Jacopo da San Cassiano

- Part of Archimedes' works was translated by William of Moerbeke (1269)
- However, given also William's very literal translation, which entailed a rather 'obscure' Latin, the task was undertaken anew in the XV century by Jacopo da San Cassiano
- Jacopo's translation circulated widely in the Renaissance and was eventually published within the *editio princeps* of Archimedes' work
- This resulted in an increased accessibility of Archimedes' text. But what were the language features that made the text more readable?
- How did Jacopo adapt (Neo)Latin linguistic features to render Greek mathematical language?
- Possible to analyze Jacopo's autograph under this point of view (Nouv. Acq. Lat. 1538)

\overline{EC} . disjungendo et permutatim argumentabimur. sicut
 \overline{Ka} ad \overline{ch} . hoc est \overline{hc} . ad \overline{cd} . ut \overline{ae} . ad \overline{ec} . Rursus
 quoniam sicut \overline{KE} ad \overline{ea} . sic utraq; simul \overline{hce} . ad \overline{ce} .
 disjungendo et permutatim sicut \overline{Ka} . ad \overline{ch} . hoc est
 ad \overline{ha} . sic \overline{ae} . ad \overline{ec} . hoc est \overline{hc} . ad \overline{cd} . et coniungendo
 equalis est aut \overline{ah} . \overline{hc} . sicut \overline{g} . \overline{kh} . ad \overline{ha} . sic \overline{hd} .
 ad \overline{dc} . et sic tota \overline{kd} ad \overline{dh} . ut \overline{dh} . ad \overline{dc} . hoc est ut
 \overline{kh} . ad \overline{ha} . quadratur \overline{g} \overline{kh} . \overline{ha} . equatur \overline{g} \overline{kh} . \overline{hd} . \overline{dc} . \overline{dh} . \overline{kc} . continet
 Rursus quoniam est sicut \overline{kh} . ad \overline{hc} . sic \overline{hd} . ad \overline{cd} . et
 permutatim: at uo sicut \overline{hc} . ad \overline{cd} . ostensum est sic \overline{ce}
 \overline{de} . ad \overline{ec} . sicut \overline{g} \overline{kh} . ad \overline{hd} . sic \overline{he} . ad \overline{ec} . quare
 et sic quadratur \overline{kd} . ad \overline{g} \overline{kh} . \overline{hd} . sic quadratur
 \overline{ac} . ad \overline{g} \overline{kh} . \overline{ec} . quadratur \overline{g} \overline{kh} . \overline{hd} . ostensum
 est equale esse \overline{kd} \overline{ah} . et qd sibi \overline{kd} \overline{ah} . continetur. sicut

Nouv. Acq. Lat. 1538, f. 42r (Jacopo's autograph)

MauroTex Intro

- MauroTex is a language for transcription and edition of texts that allows to collect any number of witnesses and mark up any amendment to the text
- It's based on Latex and it was developed for edition of scientific works of Francesco Maurolico, from which it takes its name
- There are three steps at the basis of the edition:
 1. transcribe manuscript in a *file.tex*
 2. pre-process *file.tex* with *m2lv* into a *file.m.tex*
 3. compile *file.m.tex* to get *file.m.pdf* as output
- It is possible to have an html file as output by using a different processor (*m2hv*)

MauroTex and Jacopo's transcription

[Na:103v]

linea recta quae circumducta fuit comprhensum tertiam partem esse circuli eius¹⁴ qui centrum habeat punctum quiescens intervallum vero secundum¹⁵ eam lineae motae partem quae a puncto moto fuerit in una circumvolutione permeata; et si lineam spiralem lineam recta contigerit in puncto quod fuit in spirali¹⁶ ultimo productum alia item linea recta a puncto circumductae quiescente ducatur ad ipsam circumductam et in locum unde moveri¹⁷ ceperat regressam secundum angulos rectos quousque cum contingente concurrat dico hanc lineam productam circumferentiae circuli in prima circumvolutione producti esse aequalem. **30** Item si linea circumducta¹⁸ et punctum latum secundum illam pluribus circumvolutionibus circumferantur et in locum unde moveri ceperint multotiens restituantur dico spacii illius quod in¹⁹ secunda circumvolutione fuerit | a spirali linea comprhensum, duplum illud existet quod in tertia comprhendetur. Quod vero in quarta triplum quod in quinta quadruplum et sic deinceps semper spacium in posterioribus²⁰ circumvolutionibus conclusa secundum consequens augmentum numerorum multiplicia erunt ad spacium in secunda revolutione conclusum, spacium vero in prima revolutione contentum sexta pars existet spacii in secunda revolutione comprhensi.

31 Item si in spirali linea duo puncta notentur et ab eis iungantur lineae rectae ad terminum lineae circumductae quiescentem et duo circuli describantur centro quod sit punctum quiescens secundum intervala duarum linearum rectarum quae ad quiescentem lineam spiralem terminum ductae fuerint et earum linearum minor extra ducatur dico spacium comprhensum circumferentiae maioris circuli parte illa quae in eandem partem cum linea spirali fertur mediaque²¹ inter lineam rectam²² et spiralem lineam habeantur et a linea recta extra ducta et a linea spirali²³ ad spacium comprhensum sub minoris circuli ea circumferentiae parte quae inter eandem lineam spiralem et lineam²⁴ rectam media existit et sub linea quae earum terminos iungit

¹⁴eius *supra lineam Na*

¹⁵*ante* secundum *del.* aequal *Na*

¹⁶li *supra lineam Na*

¹⁷*ante* moveri *del.* mota *Na*

¹⁸circumducta *ex* circumvoluta *Na*

¹⁹*ante* in *del.* in prima *Na*

²⁰*ante* posterioribus *del.* posteribus *Na*

²¹fertur mediaque *ex* existat quod mediam *Na*

²²lineam rectam *ex* lineas rectas *Na*

²³et a linea spirali *supra lineam Na*

²⁴lineam *ex* lineas *Na*

- Nouv. Acq. Lat. 1538 is the only Archimedean autograph by Jacopo da San Cassiano
- The manuscript is a working copy: a draft. There are a lot of mistakes, correction *in scribendo* or additional words
- In the edition each type of Jacopo's correction is recorded in apparatus
- This translation had a larger fortune: thanks to MauroTex, it will be easily possible to collate Jacopo's text with its copies and study textual variants

From MauroTeX to linguistic annotation

- From the teX file, necessary to extract the text of the manuscript
- Some specific challenges:
 - Errors in the translation that make the Latin sentence grammatically “incorrect”
 - Errors in the transcription
 - Lack of punctuation
- Current strategy: 0 intervention
 - However, this might change in the future
- Pilot: *The Spirals*

Tokenization and POS tagging

- Pie Latin LASLA+ model 0.0.6 was used for tokenization and POS tagging
 - Pie model (Manjavacas et al.) fine-tuned on Latin annotated corpus of the LASLA (Thibault Clérice)
 - Adaptation to this specific case (Thank you Thibault!)
- Great advantage: Post-correction interface Pyrrha
 - Each modification can be applied to all «similar tokens»
 - Perfectly suited for mathematical language
- No POS tag associated to mathematical letters: SYM or NOUN?

\overline{ec} . disiungendo et permutatum argumentabimur. sicut
 \overline{ka} ad \overline{ch} . hoc est \overline{hc} . ad \overline{cd} . ut \overline{ae} . ad \overline{ec} . Rursus
 quoniam sicut \overline{ke} ad \overline{ea} . sic utraq; simul \overline{hce} . ad \overline{ce} .
 disiungendo et permutatum sicut \overline{ka} . ad \overline{ch} . hoc est
 ad \overline{ha} . sic \overline{ae} . ad \overline{ec} . hoc est \overline{hc} . ad \overline{cd} . et coniungendo
 equalis est aut \overline{ah} . \overline{hc} . sicut \overline{kh} . ad \overline{ha} . sic \overline{hd} .
 ad \overline{dc} . et sic tota \overline{kd} ad \overline{dh} . ut \overline{dh} . ad \overline{dc} . hoc est ut
 \overline{kh} . ad \overline{ha} . ^{continet sub} \overline{kd} \overline{ha} . ^{ei} \overline{hd} \overline{kc} . ^{sub} \overline{dc} . ^{continet}
 Rursus quoniam est sicut \overline{kh} . ad \overline{hc} . sic \overline{hd} . ad \overline{cd} . et
 permutatum: at non sicut \overline{hc} . ad \overline{cd} . ostensum est sic \overline{ce}
 \overline{de} . ad \overline{ec} . sicut \overline{kh} . ad \overline{hd} . sic \overline{he} . ad \overline{ec} . quare
 et sic ^{id} \overline{kd} . ad ^{continet sub} \overline{hd} . sic ^{id} \overline{kd} . ad ^{continet sub} \overline{hd} . sic ^{id} \overline{kd} . ad ^{continet sub} \overline{hd} .
 \overline{ac} . ad ^{id} \overline{ec} . ^{continet sub} \overline{ae} . ^{continet sub} \overline{ec} . ^{id} \overline{kd} . ^{continet sub} \overline{hd} . ostensum
 est ^{id} \overline{kd} \overline{ah} . et qd sub \overline{kd} \overline{ah} . continetur. sicut

Nouv. Acq. Lat. 1538, f. 42r (Jacopo's autograph)

Biaffine Parser and Treebank embeddings

- Deep Biaffine Parser used to parse the text
- Implementation with the MaChamp library: treebank (dataset) embeddings
- When training with multiple (and non-homogenous) treebanks, one embedding is added that captures the features of each treebank
- When parsing a new text, a treebank identifier is given
- The parser follows the “style” of the chosen treebank
- Suited for mathematical Latin? Non-classical, highly regular language

Data creation and annotation

- After a first run of the Biaffine parser trained on cluster of ancient languages
- Universal Dependencies treebank
- As treebank embedding, we used UDante, first “native” UD Latin treebank
- Manual correction (UD Annotatrix) for ca. 1200 tokens of the *Spirals*, used then to train a second time the parser
- The parser, with the embedding of the *Spirals*, was used to parse a new portion of the *Spirals*

Some annotation problems

- Extremely long sentences: difficult to make sense of the link between the clauses (parataxis? coordination?)
 - median length is of 21.5 words with a maximum of 104,
 - syntactic trees with a median depth of 5.5 layers
- Latin particles: discourse? Coordination?
- “Linea AB”: chosen the flat relation, but the parser tends to always assign “nmod” (maybe it is right?)
- Lack of language-specific and “genre”-specific guidelines!

Evaluation

- Better than the baseline!
- But far from perfect 😞
- Biaffine parser without mathematical training data:
 - UAS: 70.56
 - LAS: 58.63

Model	POS	UAS	LAS
Biaffine Archimedes	91.25	72.43	59.85
IT-TB	NA	68.60	55.03
Perseus	NA	68.16	50.44

Table 1: UPOS, UAS and LAS score of different parsers

Where did it go wrong?

- POS:
 - *Spiralis* (quite a key-word in *The Spirals*) always tagged as noun whereas it is an adjective in the expression *linea spiralis*
 - Confusion between DET and PRON
- Difficult in assigning the right head to mathematical letters
- Syntactic relations:
 - *linea AB* (nmod) instead of (flat)

Conclusions

- More training data should help to better capture mathematical features
- Less noisy text
- Discussion of 'genre' specific choices: maybe wrong choices made in the first place?
- Adding already these training data improved the performance of the Biaffine parse
- Goal: create *Archimedes Latinus*