Collatinus : Lemmatizer and morphological analyzer for Latin texts.

Yves Ouvrard
Philippe Verkerk

Digital Classics III: Re-thinking Text Analysis May 12th 2017

Who are we ?

Yves Ouvrard
Professor of latin
(retired)

Philippe Verkerk
Physicist
Lab. PhLAM

What was initially Collatinus ?

- Prepare the wordlist associated to a text

- Reading support of latin texts

Collatinus 11

- Free and Open program (GNU GPL)
  - written in C++ (Qt 5)
  - resources as plain text files
- Stand-alone software (Mac OS X, Windows, Linux)

  http://outils.biblissima.fr/en/collatinus/
- On-line version (Collatinus 10.2)

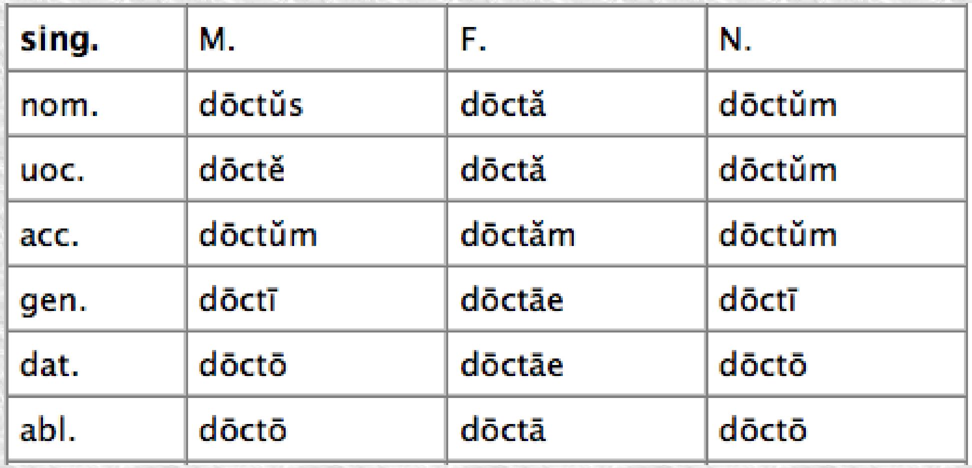
  http://outils.biblissima.fr/en/collatinus-web/

Main functions

- Lemmatization
  Analysis
  Reading support

- Dictionaries
  - Gaffiot 2016
  - Lewis & Short
  - Georges...

- Inflection



sing.	M.	F.	N.
nom.	dōctŭs	dōctă	dōctŭm
uoc.	dōctĕ	dōctă	dōctŭm
acc.	dōctŭm	dōctăm	dōctŭm
gen.	dōctī	dōctāe	dōctī
dat.	dōctō	dōctāe	dōctō
abl.	dōctō	dōctā	dōctō

- Scansion

Ārmă (Ārmā) vĭrūmquĕ cānō (cănō), Trōjāe quī prīmŭs ăb ōrīs (ōrĭs)
-̆- u-u -̆-̆ -- - -u u -̆-
Ītălĭām, fātō prŏfŭgūs, Lāvīnĭăquĕ (Lāvīnĭāquĕ) vēnĭt (vĕnĭt)
-uu- -- uu- --u-̆u -̆u
lītŏră, mūlt[um] īll[e] ēt tērrīs jāctātŭs (jāctātūs) ĕt āltō
-uu -` -` - -- --̆- u --
vī sŭpĕrūm sāevāe mĕmŏrēm Jūnōnĭs ŏb īrăm;
- uu- -- uu- --u u -u
mūltă (mūltā) quōqu[e] (quŏqu[e]) ēt bēllō pāssūs, dūm cōndĕrēt ūrbĕm,
-̆- -̆` - -- -- - -uu -u
īnfērrētquĕ dĕōs Lătĭō, gĕnūs (gĕnŭs) ūndĕ Lătīnŭm,
---u u- uu- u-̆ -u u-u
Ālbānīquĕ pātrēs (pātrēs), ātqu[e] āltāe mōenĭă Rōmāe.
---u -̆- -` -- -uu --

What makes the difference ?

- Dictionaries :
  a single click opens the dictionary
- Allows to compare two dictionaries

- Scansion :
  knows a priori if a vowel is short or long
  → not limited to metrical verses

- Integrated server
  → can answer a question from another program
- Expandable (one can add lemmas and/or dictionaries)

Principle of the analysis

- Not a list of inflected forms
  → closer to the human mechanism

- Split the word in 2 parts (all possibilities)
- Look for the 2 parts in the proper list
- Check for the compatibility

- 24 100 lemmas in the “main lexicon”
- 58 000 lemmas in the “extended lexicon”

The figures

- 24 100 lemmas (classical Latin)
- 58 000 lemmas in the extended lexicon (unusual words)
- 140 “paradigms”

- Lewis & Short ≈ 60 000 entries
- Gaffiot 2016 ≈ 60 000 entries
- K.E. Georges ≈ 50 000 entries
- Gérard Jeanneau ≈ 50 000 entries

Generate automatically the lemmas in the format of Collatinus

7 500 entries still waiting

- Assimilation of the prefix (without an extra entry)

  obfero ↔ *offero*

- Contraction in the declension

  ĭĭ → *ī*

  amasse ↔ *amavisse*

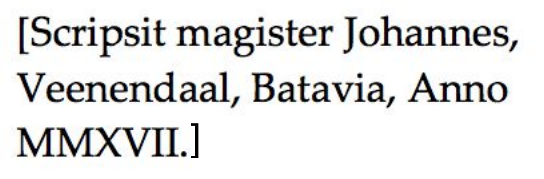
Ordering the analysis

- “Ordering” the analysis
  → a step to desambiguization
  ≠ from a usual POS tagger (no list of forms)

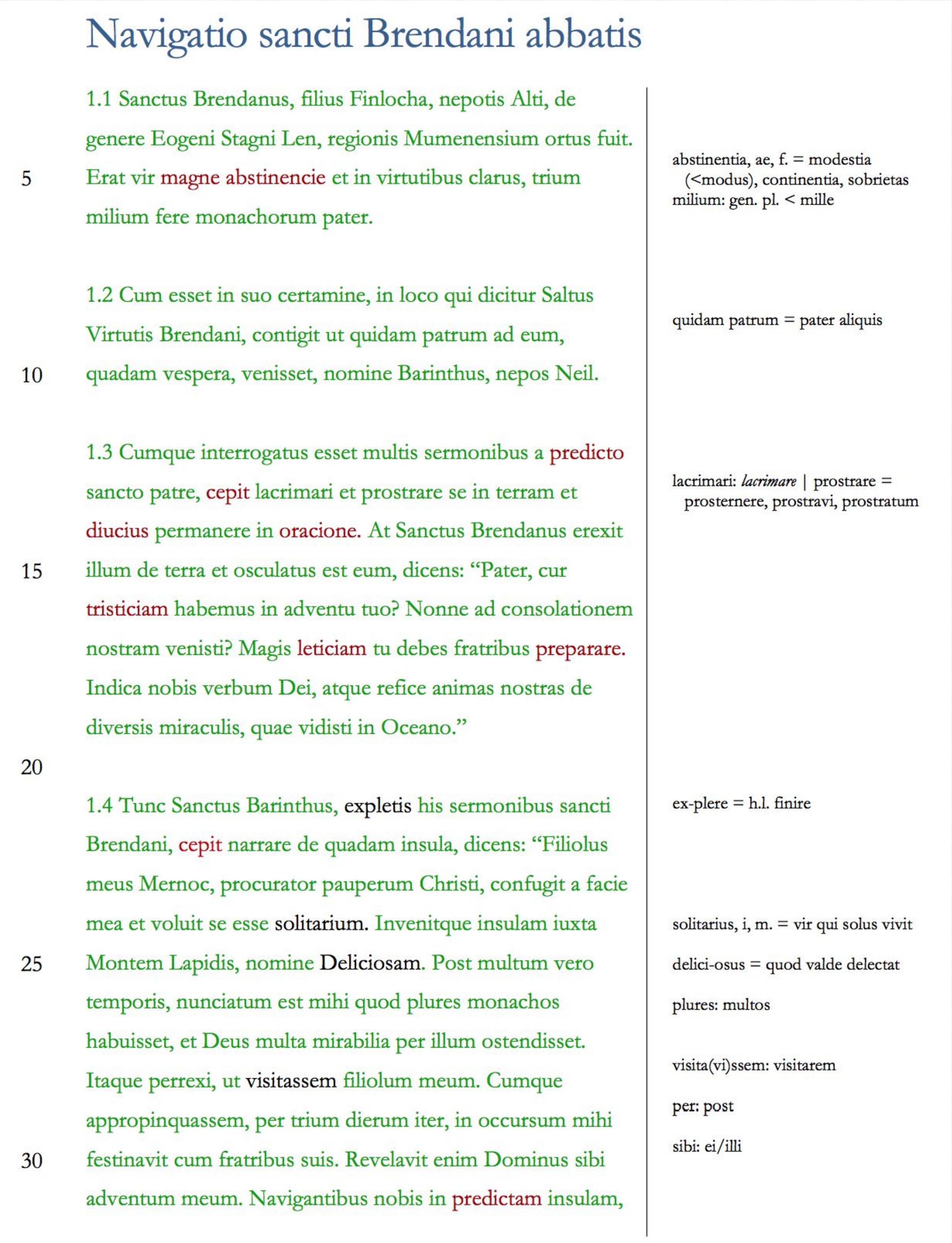
- Statistics on the texts lemmatized by the LASLA (almost 2 000 000 forms)
- Counting the tags, the sequences of 3 tags and each lemma (mapped in the lexicon)
- 91 tags : more than the POS

TextiColor

- Students are supposed to know a list of words
- The text is prepared with different colors :
  - known
  - not known
  - medieval form



[Scripsit magister Johannes, Veenendaal, Batavia, Anno MMXVII.]



Navigatio sancti Brendani abbatis

1.1 Sanctus Brendanus, filius Finlocha, nepotis Alti, de genere Eogeni Stagni Len, regionis Mumenensium ortus fuit. Erat vir magne abstinencie et in virtutibus clarus, trium milium fere monachorum pater.

1.2 Cum esset in suo certamine, in loco qui dicitur Saltus Virtutis Brendani, contigit ut quidam patrum ad eum, quadam vespera, venisset, nomine Barinthus, nepos Neil.

1.3 Cumque interrogatus esset multis sermonibus a predicto sancto patre, cepit lacrimari et prostrare se in terram et diucius permanere in oracione. At Sanctus Brendanus erexit illum de terra et osculatus est eum, dicens: “Pater, cur tristiciam habemus in adventu tuo? Nonne ad consolationem nostram venisti? Magis leticiam tu debes fratribus preparare. Indica nobis verbum Dei, atque refice animas nostras de diversis miraculis, quae vidisti in Oceano.”

1.4 Tunc Sanctus Barinthus, expletis his sermonibus sancti Brendani, cepit narrare de quadam insula, dicens: “Filiolus meus Mernoc, procurator pauperum Christi, confugit a facie mea et voluit se esse solitarium. Invenitque insulam iuxta Montem Lapidis, nomine Deliciosam. Post multum vero temporis, nunciatum est mihi quod plures monachos habuisset, et Deus multa mirabilia per illum ostendisset. Itaque perrexi, ut visitassem filiolum meum. Cumque appropinquassem, per trium dierum iter, in occursum mihi festinavit cum fratribus suis. Revelavit enim Dominus sibi adventum meum. Navigantibus nobis in predictam insulam,

abstinentia, ae, f. = modestia (<modus), continentia, sobrietas
milium: gen. pl. < mille

quidam patrum = pater aliquis

lacrimari: *lacrimare* | prostrare = prosternere, prostravi, prostratum

ex-plere = h.l. finire

solitarius, i, m. = vir qui solus vivit

delici-osus = quod valde delectat

plures: multos

visita(vi)ssem: visitarem

per: post

sibi: ei/illi

From quantities to rhythm

- Collatinus knows the quantity of the vowels
  → easy to determine the tonical accent.

- Can split the syllabes → hyphenation
  abs·cí·di (abs-cido) ≠ áb·sci·di (ab-scindo)

- Opens the way to the study of rhythmic prose
- Statistics on paroxytons and proparoxytons.
- Rhythm everywhere, not only in *clausulae*

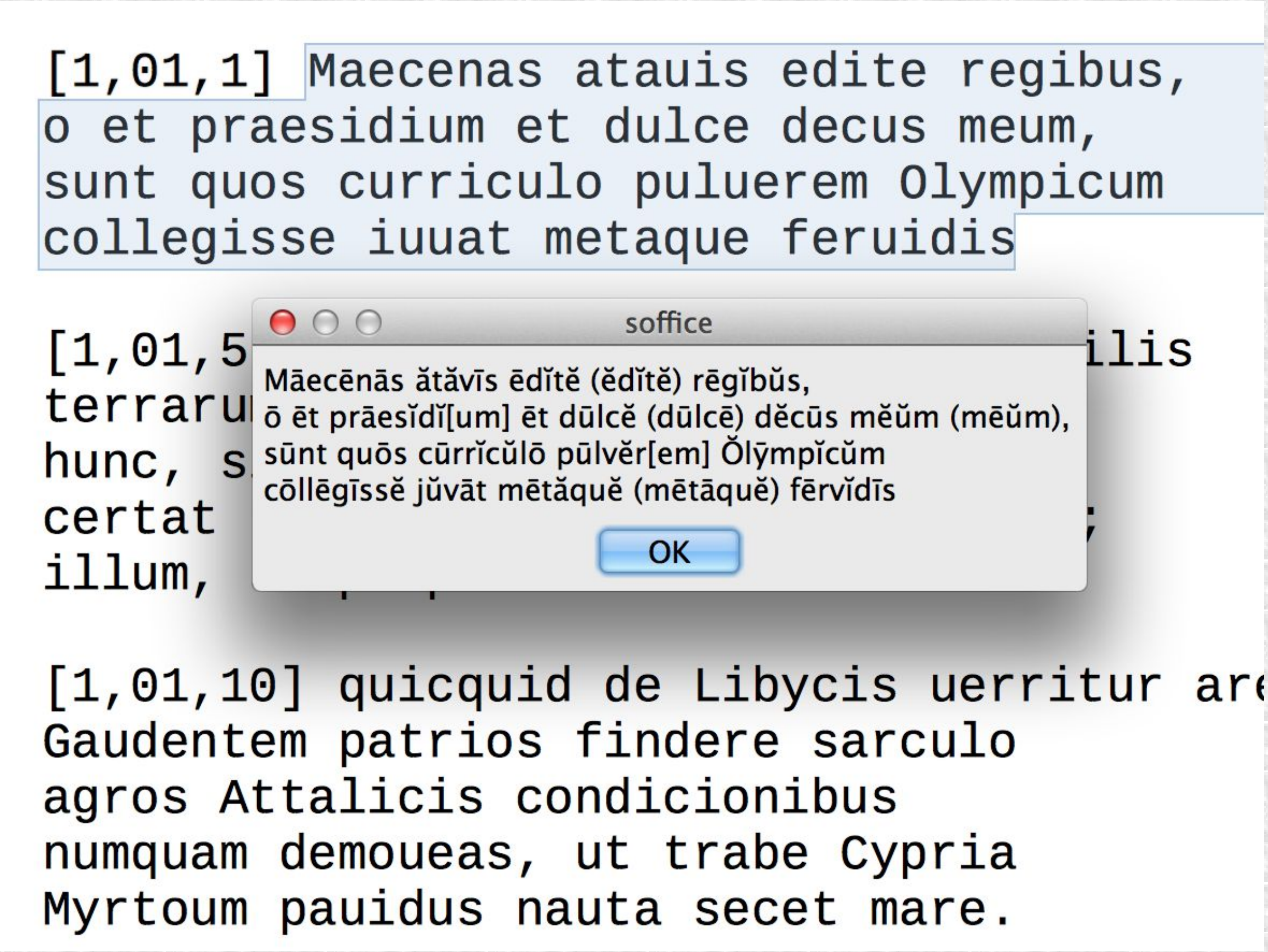
New feature : probabilistic tagger

- 2nd order hidden Markov model
- Training corpus : LASLA.
- For each sentence, chooses the best sequence of tags and thus the best analysis

- No one knows if a probabilistic tagger works with Latin (“free” order of the words)
- No evaluation of the error rate (not yet)

Internal server

- Collatinus can answer questions asked from another program. For instance, LibreOffice.



[1,01,1] Maecenas atauis edite regibus,
o et praesidium et dulce decus meum,
sunt quos curriculo puluerem Olympicum
collegisse iuuat metaque feruidis

[1,01,5 ilis
terraru
hunc, s
certat ;
illum,

[1,01,10] quicquid de Libycis uerritur are
Gaudentem patrios findere sarculo
agros Attalicis condicionibus
numquam demoueas, ut trabe Cypria
Myrtoum pauidus nauta secet mare.

Māecēnās ătăvīs ēdĭtě (ĕdĭtě) rēgĭbŭs,
ō ēt prāesĭdĭ[um] ēt dūlcĕ (dūlcē) dĕcūs mĕŭm (mēŭm),
sūnt quōs cūrrĭcŭlō pūlvĕr[em] Ŏlȳmpĭcŭm
cōllēgīssĕ jŭvāt mētăquĕ (mētāquĕ) fērvĭdīs

Internal server (2)

- A “standard” tagger with the data from LASLA
  - List of forms, with lemma and analysis
  - Exact number of occurrences
  - Tagset and statistics on sequences of tags

- Problem of the forms that have not been seen before : Ask Collatinus !

  → Answers a LASLA-like analysis

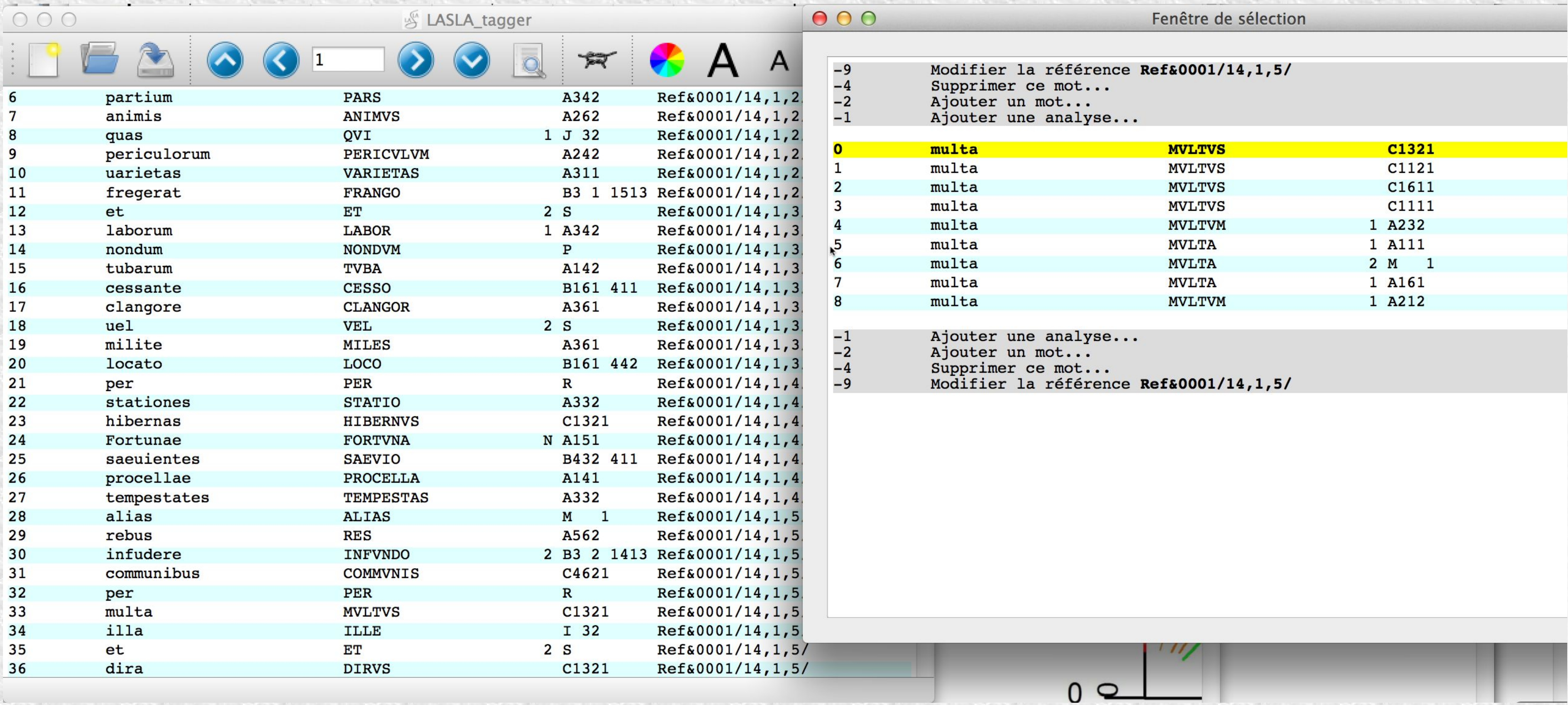
LASLA-Tagger

- Supervised lemmatization.
- Edition/correction of the result of the tagger.

→ Hope to reach zero error !

- Exact number of occurrences. For “patres” : 253 voc. 110 nom. and 75 acc.
  Collatinus expects 6 voc. 200 nom. & 320 acc

LASLA-Tagger : screen-shot



```
6    partium       PARS              A342      Ref&0001/14,1,2
7    animis        ANIMVS            A262      Ref&0001/14,1,2
8    quas          QVI             1 J 32      Ref&0001/14,1,2
9    periculorum   PERICVLVM         A242      Ref&0001/14,1,2
10   uarietas      VARIETAS          A311      Ref&0001/14,1,2
11   fregerat      FRANGO            B3 1 1513 Ref&0001/14,1,2
12   et            ET              2 S         Ref&0001/14,1,3
13   laborum       LABOR           1 A342      Ref&0001/14,1,3
14   nondum        NONDVM            P         Ref&0001/14,1,3
15   tubarum       TVBA              A142      Ref&0001/14,1,3
16   cessante      CESSO             B161 411  Ref&0001/14,1,3
17   clangore      CLANGOR           A361      Ref&0001/14,1,3
18   uel           VEL             2 S         Ref&0001/14,1,3
19   milite        MILES             A361      Ref&0001/14,1,3
20   locato        LOCO              B161 442  Ref&0001/14,1,3
21   per           PER               R         Ref&0001/14,1,4
22   stationes     STATIO            A332      Ref&0001/14,1,4
23   hibernas      HIBERNVS          C1321     Ref&0001/14,1,4
24   Fortunae      FORTVNA         N A151      Ref&0001/14,1,4
25   saeuientes    SAEVIO            B432 411  Ref&0001/14,1,4
26   procellae     PROCELLA          A141      Ref&0001/14,1,4
27   tempestates   TEMPESTAS         A332      Ref&0001/14,1,4
28   alias         ALIAS             M   1     Ref&0001/14,1,5
29   rebus         RES               A562      Ref&0001/14,1,5
30   infudere      INFVNDO         2 B3 2 1413 Ref&0001/14,1,5
31   communibus    COMMVNIS          C4621     Ref&0001/14,1,5
32   per           PER               R         Ref&0001/14,1,5
33   multa         MVLTVS            C1321     Ref&0001/14,1,5
34   illa          ILLE              I 32      Ref&0001/14,1,5
35   et            ET              2 S         Ref&0001/14,1,5/
36   dira          DIRVS             C1321     Ref&0001/14,1,5/
```

```
-9        Modifier la référence Ref&0001/14,1,5/
-4        Supprimer ce mot...
-2        Ajouter un mot...
-1        Ajouter une analyse...

0         multa                 MVLTVS              C1321
1         multa                 MVLTVS              C1121
2         multa                 MVLTVS              C1611
3         multa                 MVLTVS              C1111
4         multa                 MVLTVM            1 A232
5         multa                 MVLTA             1 A111
6         multa                 MVLTA             2 M    1
7         multa                 MVLTA             1 A161
8         multa                 MVLTVM            1 A212

-1        Ajouter une analyse...
-2        Ajouter un mot...
-4        Supprimer ce mot...
-9        Modifier la référence Ref&0001/14,1,5/
```

What is going on ?

***Praelector*:**

- “Grammatical assistant” to help pupils in reading Latin / understanding the sentence.

  → “Translation” in French of the sentence

- Medieval spelling :
  e for ae, o for au, f for ph, etc...
  Reduce the medieval forms and the classical words to the same “phonetic” form.

Future plans

- Coupling of the probabilistic tagger with the gramatical analysis of the sentence
- Statistics on rhythmic prose (cursus)
- Looking for verses in prose
- “Universal” output format to gather all the information about the text (lemmas, analysis, scanned and accented forms...)

Suggestions

- Yves.Ouvrard@collatinus.org
- Philippe.Verkerk@univ-lille1.fr

Thank you for your attention

- http://outils.biblissima.fr/en/collatinus/
- http://outils.biblissima.fr/en/collatinus-web/
- For ancient Greek : http://outils.biblissima.fr/en/eulexis/