



COLLATINUS : A NEW TOOL FOR LATINISTS.

Yves Ouvrard
Philippe Verkerk

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 11.2

- Free and Open program (GNU GPL)
 - written in C++ (Qt 5)
 - sources on GitHub
 - resources as plain text files
- Stand-alone software (Mac OS X, Windows, Linux)
<http://outils.biblissima.fr/en/collatinus/>
- On-line version
<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	dōctā	dōctŭm
uoc.	dōctĕ	dōctā	dōctŭm
acc.	dōctŭm	dōctām	dōctŭm
gen.	dōctī	dōctāe	dōctī
dat.	dōctō	dōctāe	dōctō
abl.	dōctō	dōctā	dōctō

- Scansion

Ārmā (Ārmā) vīrŭmquĕ cānō (cānō), Trōjāe quī prīmŭs āb ōrīs (ōrīs)
 -- ŭ-u -- -- -- -u u --
 Ītālīām, fātō prōfŭgŭs, Lāvīniāquĕ (Lāvīniāquĕ) vēnīt (vēnīt)
 -uu- -- uu- --u-ŭ -ŭ
 Ītōrā, mŭlt[um] ĭll[e] ēt tērrīs jāctātŭs (jāctātŭs) ēt āltō
 -uu - - - - -u --
 vī sŭpĕrŭm sāevāe mēmōrēm Jūnōnīs ōb ĭrām;
 - uu- -- uu- --u u -u
 mŭltā (mŭltā) quōqu[e] (quōqu[e]) ēt bĕllō pāssŭs, dŭm cōndĕrĕt ūrbĕm,
 -- - - - -uu -u
 ĭnfĕrrĕtquĕ dĕōs Lātīō, gĕnŭs (gĕnŭs) ūndĕ Lātīnŭm,
 ---u u- uu- u-ŭ -u u-u
 Ālbānīquĕ pātrĕs (pātrĕs), ātqu[e] āltāe mōenīā Rōmāe.
 ---u -- - - -uu --

What makes the difference ?

- Dictionaries :
a single click opens
one or two dictionaries
- Allows to compare two
dictionaries
- Scansion :
knows a priori if a
vowel is short or long
- From quantities to
accented Latin.
- 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

A program has generated
automatically the lemmas in the
format of Collatinus

7 500 entries still waiting

- Assimilation of the prefix
(without an extra entry)
obfero ↔ *offero*

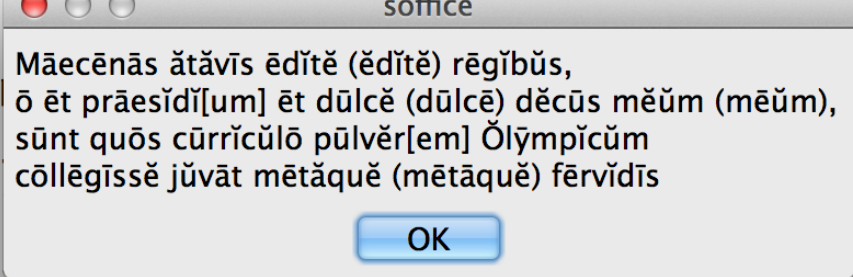
- Contraction in the declension
 $\text{ī} \rightarrow \text{ī}$
amasse ↔ *amavisse*

Internal server

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

```
[1,01,1] Maecenas atavis edite regibus,  
o et praesidium et dulce decus meum,  
sunt quos curriculo pulverem Olympicum  
collegisse iuvat metaque feruidis
```

```
[1,01,5] terrarum  
hunc, s  
certat  
illum,
```



Māecēnās ātāvīs ēdītē (ēdītē) rēgībūs,
ō ēt prāesīdī[um] ēt dūlcē (dūlcē) dēcūs mēūm (mēūm),
sūnt quōs cūrrīcūlō pūlvēr[em] Ōlūmpīcūm
cōllēgīssē jūvāt mētāquē (mētāquē) fērūidīs

OK

```
[1,01,10] quicquid de Libycis uerritur arc  
Gaudentem patrios findere sarculo  
agros Attalicas condicionibus  
numquam demoueas, ut trabe Cypria  
Myrtoum pavidus nauta secet mare.
```


Collatinus' projects

- Towards Medieval Latin
 - handle the main variations in graphy
 - expand the lexicon (Du Cange, with the help of LemLat3; toponyms, patronyms...)
- Collatinus 12 : a modular approach to Latin
- LASLA-Tagger
- Praelector (reading assistant)
- Build automatically the syntactic tree

Toward Medieval Latin

- A first attempt to handle 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.
- Du Cange \approx 90 000 lemmas
- Toponyms \approx 50 000 names
- Patronyms ?

Collatinus 12 : a modular approach

- Constant evolution of the Latin language :
From earliest latin, c. 254 BC ...
to latest medieval latin : end of XVth century.
- Medieval Latin is a natural evolution :
 - New words (and some not in use anymore)
 - New spellings or inflexions
 - New meanings

Evolutions

- New words :
feudum, i, n. : fief
garcio, onis, m. : boy (fr. garçon)
clocca, ae, f. : bell (fr. cloche)
- Inflexion :
Plautus: turrem; class. turrim
Plautus: amarier; class. amari
class.: cognitum; med.
cognoscitum
- Spelling :
Plautus: ameicus; class.
amicus
class.: aperio; med. apperio
class.: mihi; med. michi
med.: æ, œ, ã, ae->e, etc.
- Semantics :
coactio,
class. collecting, abridgment
med. applied force, necessity
curia,
class. a division of the Roman
people, place of its assembly
med. the king's court; law court

Modules and packages

- Users can load modules fitting their text.
- Chosen modules are loaded first, before classical data. So if a word is already known by a module, it is ignored.
- Modules will be compressed into packages, downloadable online.
- Collatinus 12 can uncompress and install them.

Building and using modules

- Collatinus data is readable, syntax is complex
- ECCE (Ecce Collatinistarum Communitatis Editor) is a graphical interface that will help building modules
- Better lemmatization with the right module
- Identification ? Lemmatization with different modules, comparison of scores.

LASLA-Tagger

- Need to reach zero error !
- A new text contains new forms.
 - Ask Collatinus !
Who will give a LASLA-like answer
- Supervised lemmatization.
- Edition/correction of the result of the tagger.
- 4 texts (26000 words) \approx 120 words/hour

LASLA-Tagger : screen-shot

LASLA_tagger

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	INFVNDQ	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/

Fenêtre de sélection

-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/		

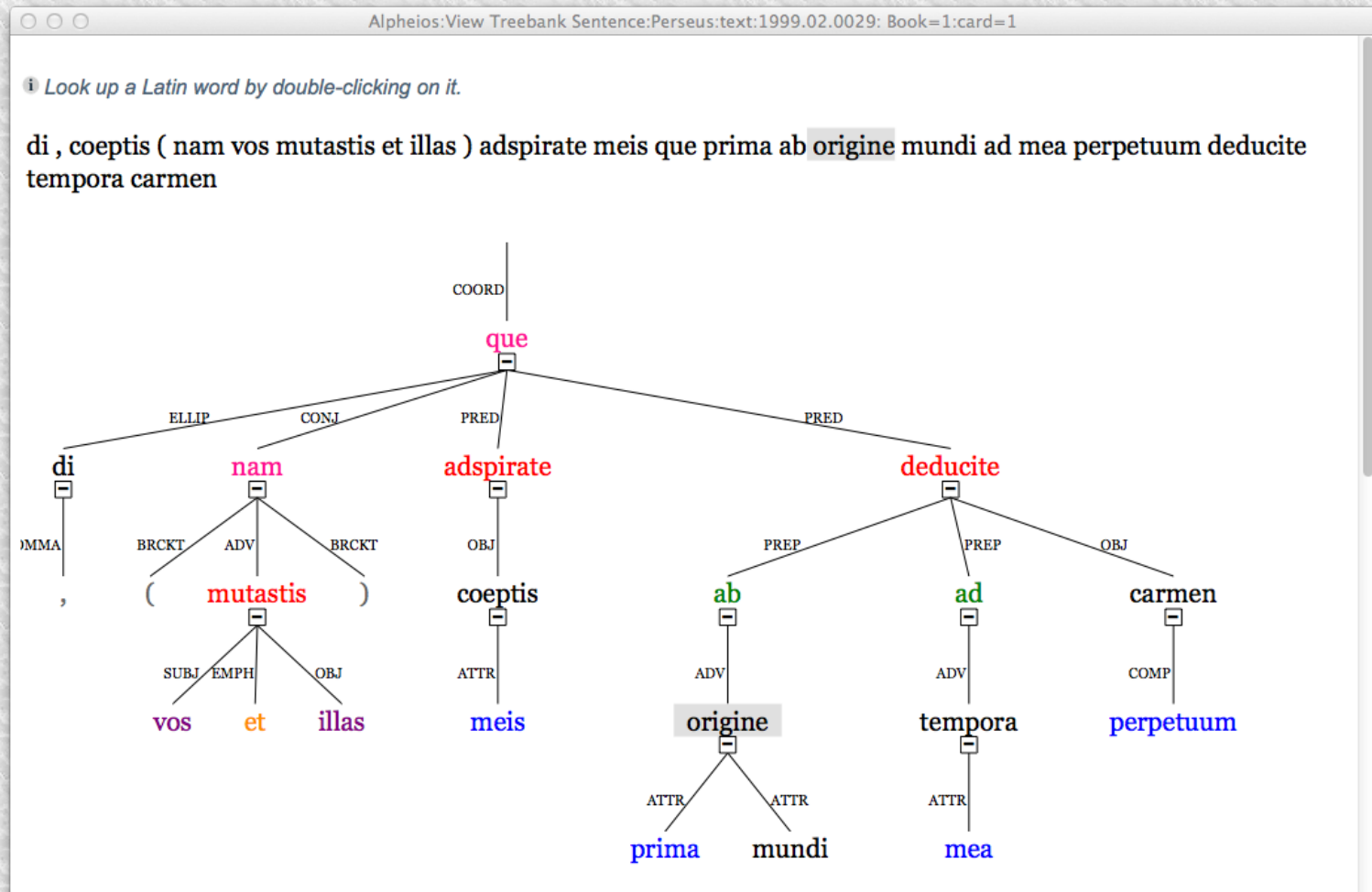
0 2

Praelector

- Reading assistant
For each pair of words, the program suggests the possible syntactic links and a translation in French, for the moment.
- The pieces of translation are put together.
- The reader can edit the translation of the sentence.

Dream : grammar

- (semi)automatic construction of syntactic trees



First steps (easy)

- Lemmatization of each word
- Order the possible solutions :
 - the result of the tagger
- Set of syntactic rules (Yves' Praelector)
- For each pair of words, try to link the various possible analyses
- Order the possible links (criteria ?).

Last steps (difficult)

- Try all the combinations of all the possible links to build a complete tree.
- Order the possible solutions (criteria ?)
- Display the “best” tree.

- Edit the “best” tree
 - Eliminate/Validate/Change some links
 - Search again (until convergence).

Difficult ?

- The number of combinations grows exponentially with the length of the sentence
- Latin language can omit a word
- Projectivity not always valid

- In most cases, one and only one link arrives on a word (except relative pronouns and predicate)
- The choice of a link imposes the analyses of two words, which eliminate some other links.

Order ? Criteria ?

- Order of syntactic links taking into account
 - the order of the analyses
 - the “frequency” of the link
- In the construction of the trees, try the “most probable” links first
- Ordering of the trees
 - Aesthetics, length...
 - Number of orphans, of link-crossing...

Thank you for your attention !