

The Clotho Project



Thibault Clérice
ICS London seminars: Summer 2014

Source of the problem :

Understand and translate *lascivus, a, um*

- Dictionary definitions really « light » for sometimes highly licentious usages
- Auto-censorship ?
- Usage quite often in no-doubt sexual contexts
- ... But also non sexual contexts
- Michel Dubuisson « Lasciva Venus, small guide to Roman Love »

Source of the problem :

Understand and translate *lascivus, a, um*

« id iudicium Auli filio condonatum putabatur; et Epicratem suspicor, ut scribis, lascivum fuisse. »

Cicéron, Lettres à Atticus, II, 3 :

« On admettait l'attribution de cette décision au fils d'Aulus, mais je suspecte Epicrate (surnom de Pompée), comme tu l'as écrit, d'avoir fait des siennes (*lascivum*) ».

« [...] hinc nova proles / artubus infirmis teneras lasciva per herbas / ludit lacte mero mentes percussa novellas. »

Lucrèce, De rerum natura, I, 260

« [...] Voilà que les tous petits, aux membres sans force, l'esprit joyeux (*lasciva*), folâtraient dans les herbes tendres, leurs jeunes têtes chamboulées par le lait pur. »

« Lusistis, satis est : lascivi nubite cunni: Permissa est vobis non nisi casta Venus. »

Martial, Epigrammata, VI.45

« Vous avez pris du plaisir, il suffit: épousez, vagins lascifs ! On ne vous permet plus que de chastes amours »

A problem ? A context.

- An MA Dissertation
- Renewal of Ancient Studies
 - Annales and anthropological approaches
 - Gender studies and a new way to study sexuality
- Technical renewal in Humanities
 - Sources Digitization
 - Global analysis / Distant Reading
- Renewal of Linguistic
 - Stereotypical linguistic (Good bye Saussure ?)
 - Computational Linguistic
 - Contextual Meaning research

Zellig Harris et l'hypothèse distributionnelle :

“the most precise way of determining a word’s meaning is by investigating the meanings of the words that occur along with that word.”

“*How Words Carry Meaning.*” Language and Information: The Bampton Lectures, Columbia University, 1986. Lecture.

An idea :

Using or creating tools for that

- *sentiment analysis* ?
 - Find the polarity of a word in a certain context...
 - ... but not in Latin
 - Bias in our contact with Latin culture
 - We can't understand Latin through our own
 - Only the context will tell
- Alternatives ?

An idea :

Using or creating tools for that

- *Sentiment analysis via exempla*
 - Polarity indicators
 - Cultural properties indicator
 - Gender
 - Social Class
 - Geographical Origin
 - Etc.
- Context analysis
 - Sentence ?

An idea:

This is not Social Network Analysis

- Social Network Analysis
- Cultural Network Analysis
 - Network of words and names
 - Exempla
 - Locations
 - Events
 - Etc..
 - Specific relationship formats
 - Three categories : Positive, Neutral, Negative
 - Refine them with numeric weighted value in [-1;+1]



Towards a Semantic/Literature/... analysis tool

- Identify possibilities
 - No tool to provide every functions I need
 - No existing tools ?
 - Lack of communication
 - Identification of existing tool in the list of operations
 - Lemmatisation
 - NLTK
 - Gephi
- Connect it, develop the rest

Difficulties

- Automatic Lemmatisation
- Recognizing *exempla* and annotating them:
 - Rare / poor (quantitatively) DB
 - Qualitatively poor DB
- No polarity dictionary
- Annotating alone a big corpus might be difficult...

Automatic lemmatisation : Probability cleaning

$L[a;b]$ = Form (a) → Lemma (b)'s Link

$$L[a,b]=1$$

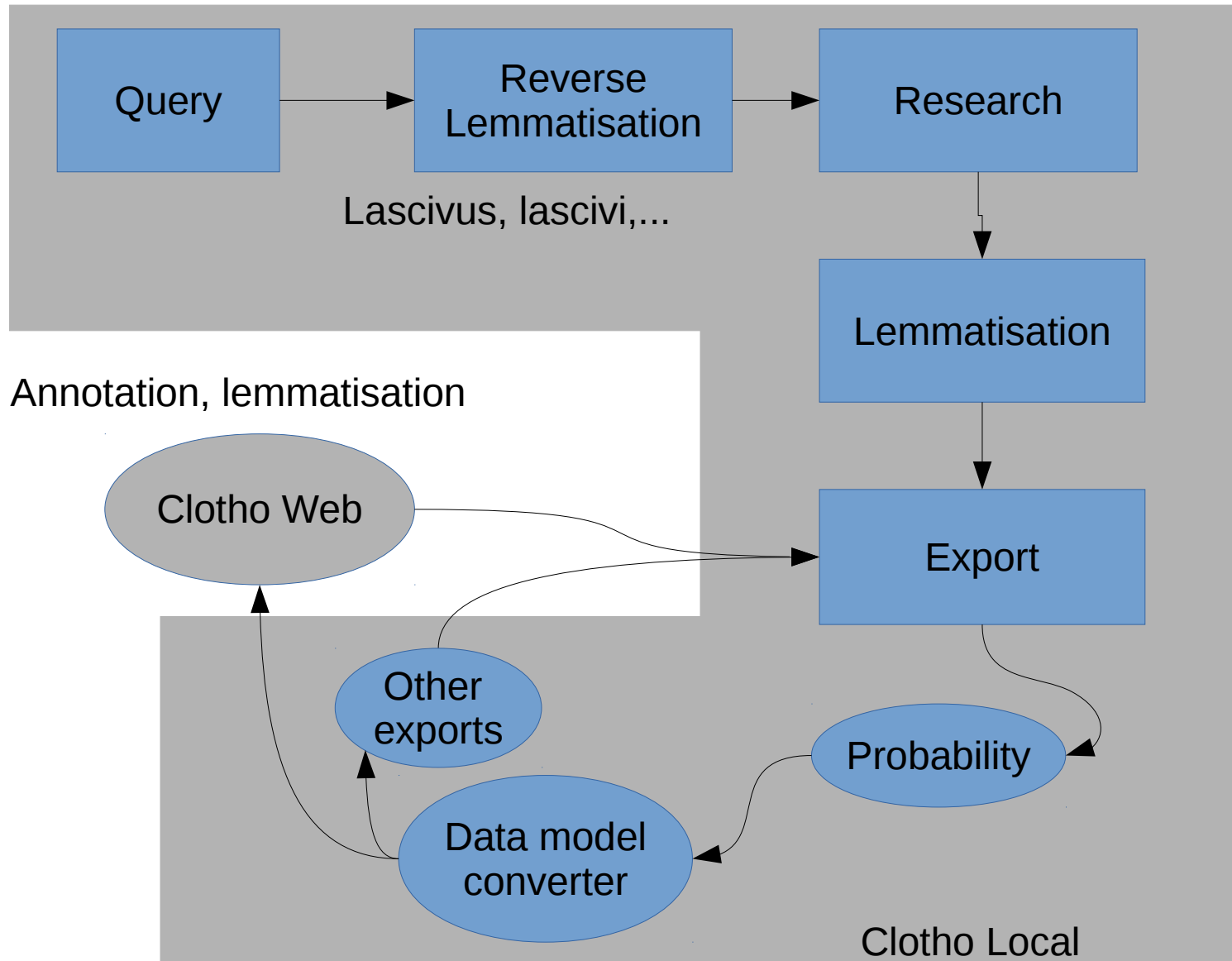
$W(l)$ = Lemma weight

$$W(l)=\sum L[f_x,l]$$

$P(L)$ = Link's probability

$$P(L[f_a:l_a])=\frac{W(l_a)}{\sum (L[f_a,l_x]\times W(l_x))}$$

Structure : The Clotho Pipeline



Crowdsourcing The interface

Aeneid

P. Vergilius Maro

[See on Perseus](#)

Poscant acies et foedera rumpant."

Haec ubi **dicta** dedit, solio se tollit ab alto
et primum **Herculeis** **sopitas** ignibus **aras**
excitat hesternumque larem parvosque **penatis**
laetus adit: mactant lectas de more bidentis
Euandrus pariter, pariter **Troiana** iuventus.

0

actions made by you here !



You are only a student here. Maybe we should do something about that... Like a vote or an annotation ?

Next Sentence

SENTENCE'S ANNOTATIONS

Polyphony	▼	Irony	▼	+
-----------	---	-------	---	---

HERCULEUS

Herculeus	1	0
New Lemma (Nominative)		Save
Mark as False positive		

POLARITY

Herculeus	<input type="radio"/> -1	<input type="radio"/> -0.5	<input type="radio"/> 0	<input type="radio"/> 0.5	<input type="radio"/> 1
-----------	--------------------------	----------------------------	-------------------------	---------------------------	-------------------------

ANNOTATION FOR HERCULEUS

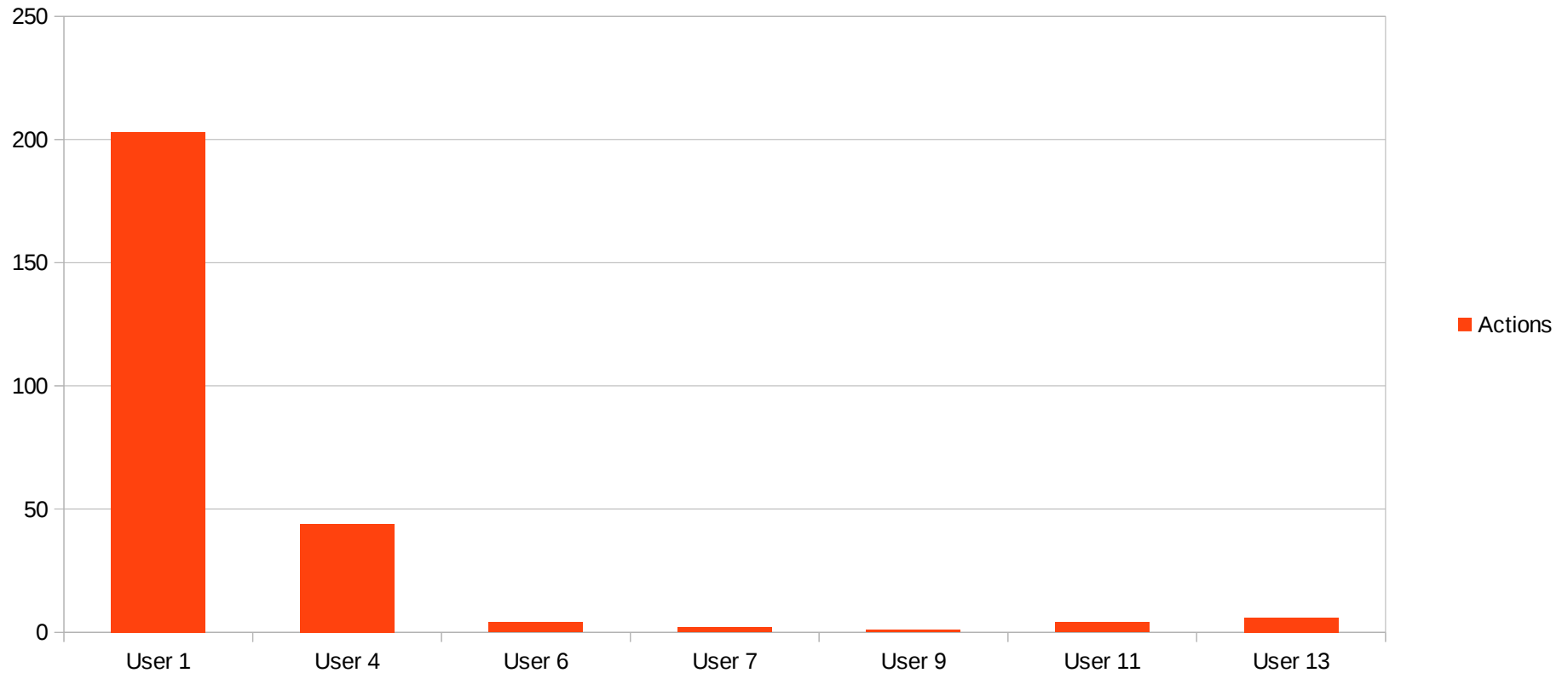
Type	▼	Date	▼	+
------	---	------	---	---

RELATIONSHIPS

Opposed to	▼	sopitas	▼	+
------------	---	---------	---	---

Crowdsourcing

Nobody likes it or understands it



264 actions, 247 including me and my tutor = 2% what is needed

Crowdsourcing Interface

Aeneid

P. Vergilius Maro

[See on Perseus](#)

Poscant acies et foedera rumpant."

Haec ubi **dicta** dedit, solio se tollit ab alto
et primum **Herculeis** **sopitas** ignibus **aras**
excitat hesternumque larem parvosque **penatis**
laetus adit: mactant lectas de more bidentis
Euandrus pariter, pariter **Troiana** iuventus.



actions made by you here !



You are only a student here. Maybe we should do something about that... Like a vote or an annotation ?

Next Sentence

SENTENCE'S ANNOTATIONS

Polyphony	▼	Irony	▼	+
-----------	---	-------	---	---

HERCULEIS

Herculeus	1	0
New Lemma (Nominative)		Save
Mark as False positive		

POLARITY

Herculeus	<input type="radio"/> -1	<input type="radio"/> -0.5	<input type="radio"/> 0	<input type="radio"/> 0.5	<input type="radio"/> 1
-----------	--------------------------	----------------------------	-------------------------	---------------------------	-------------------------

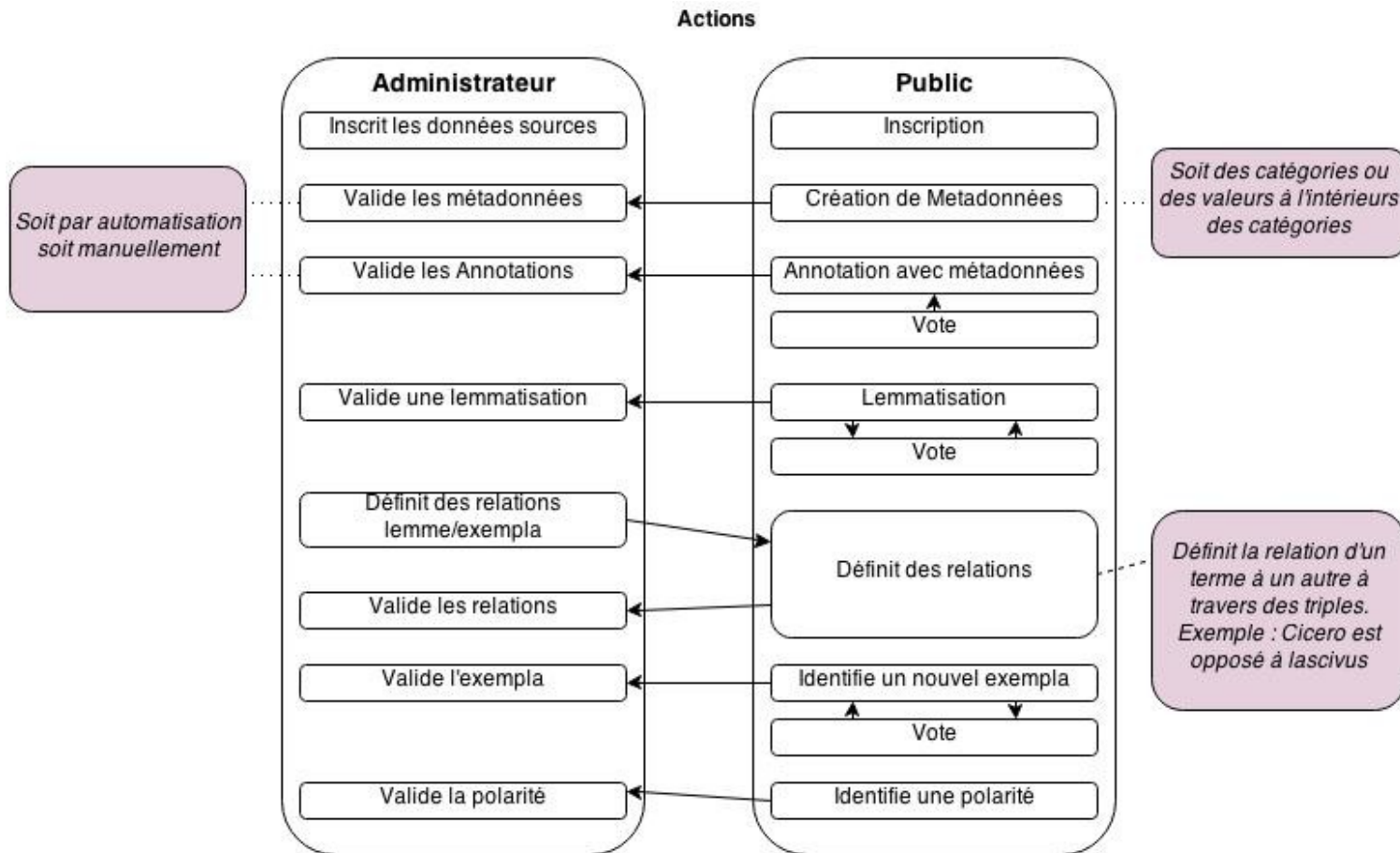
ANNOTATION FOR HERCULEUS

Type	▼	Date	▼	+
------	---	------	---	---

RELATIONSHIPS

Opposed to	▼	sopitas	▼	+
------------	---	---------	---	---

Crowdsourcing Workflow



Crowdsourcing

New idea for second interface

Participez à un grand projet, défiez vos amis sur leur connaissance en culture latine

[Connexion avec Facebook, OpenId, etc.](#)

Deuxième question du jour:

Dans la culture Romaine, Cicero est un



Vous êtes un Rhéteur ! Félicitations !

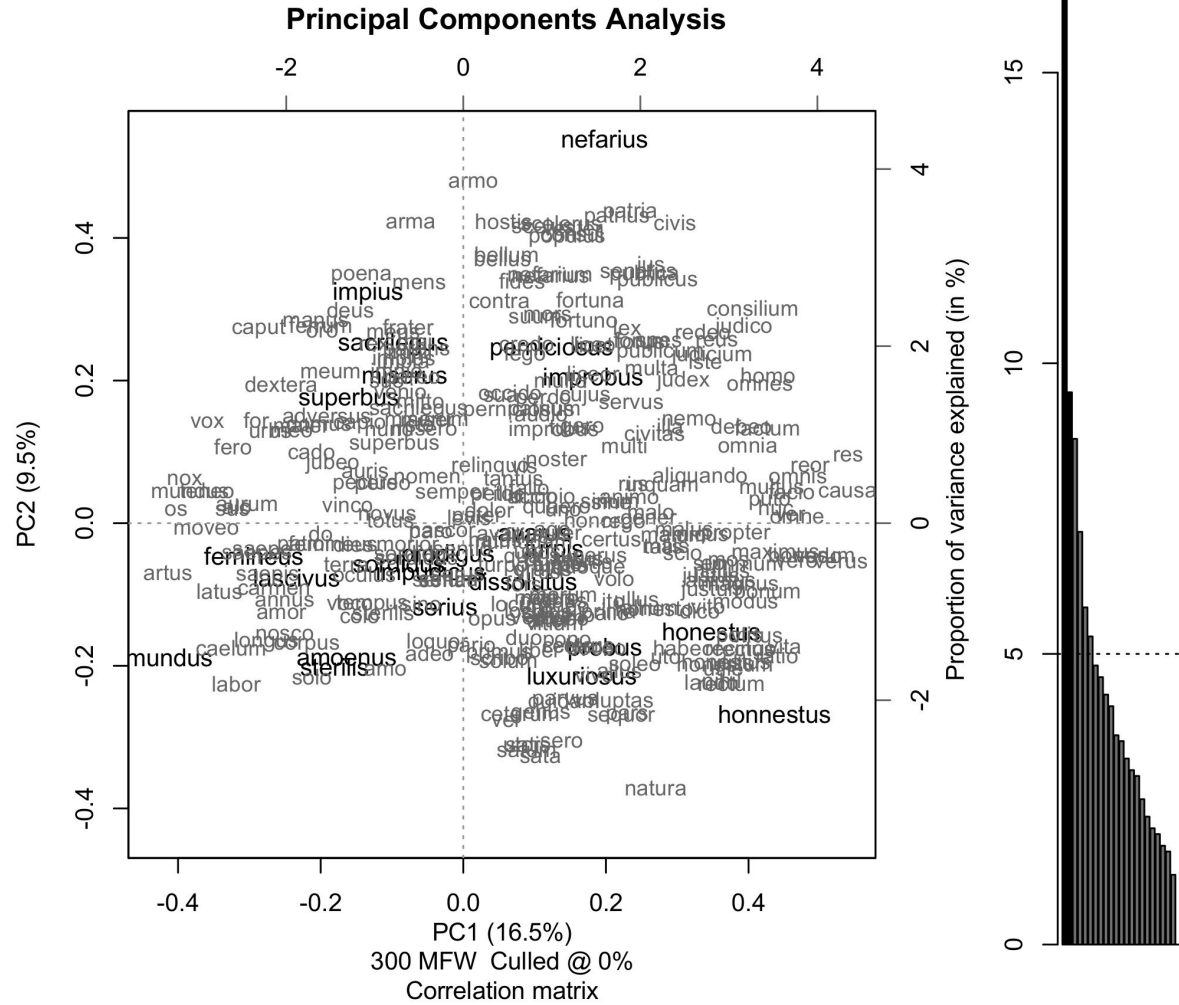
[Partager](#)



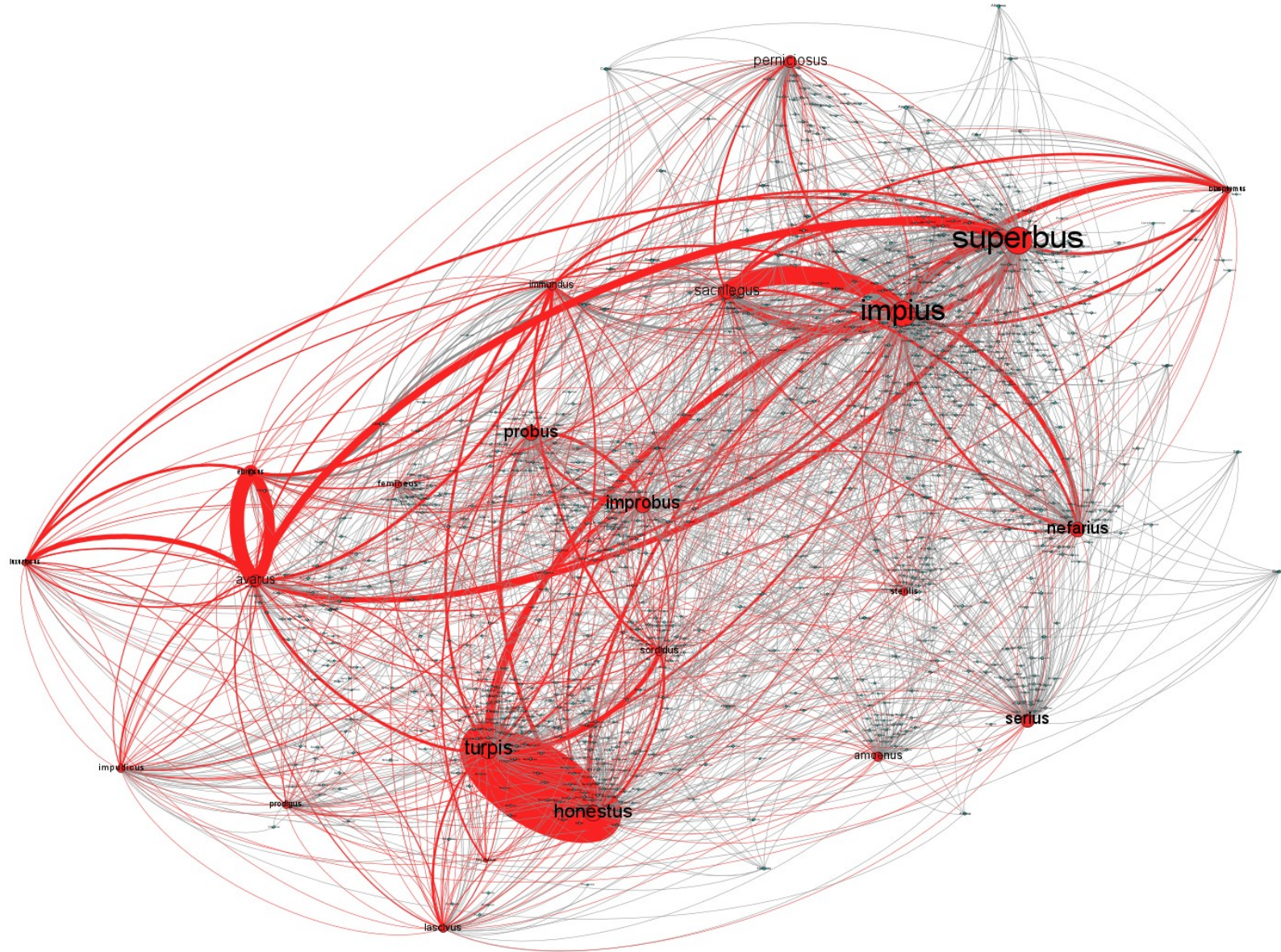
Analysis results and possibilities

- Digital analysis tend to approve traditional one
- Digital outputs ?

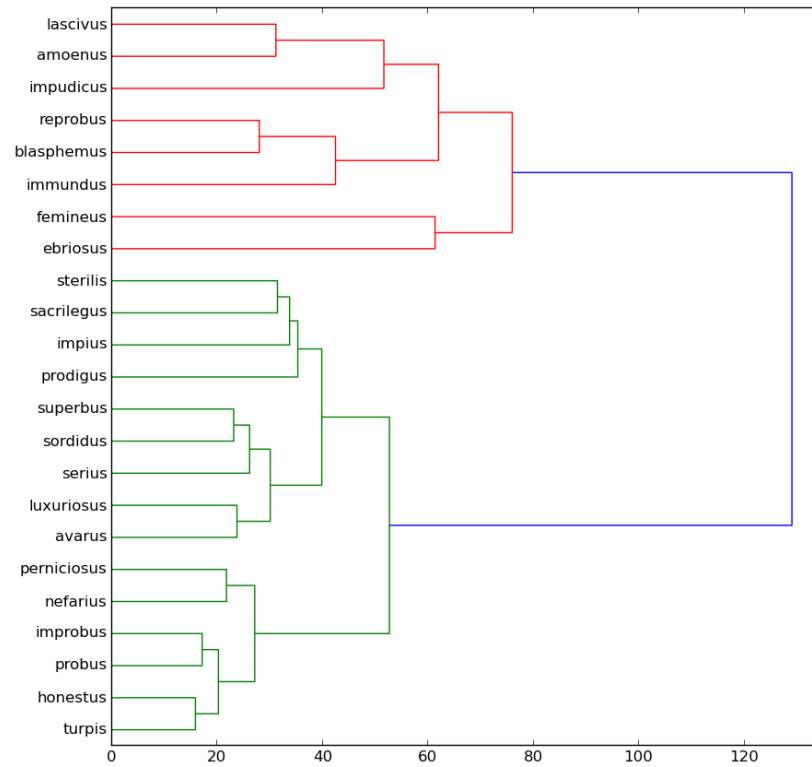
Stylo / R



Network



Dendrograms



Browser interactive exports

In the future

- Optimization
 - 760 000 / 20h
 - Cache system
 - Avoid MySQL and Lucene ?
- Lemmatisation
- Names recognition enhancement (*Marcus. T. Cicero*)
- New interface
- Network Analysis : Better annotations ?

Futur : new modules

- Context choice
 - N-words window
 - *Tree Tagging weighting*
 - *N-Gramm*
- New algorithms
 - Log-likelihood Ratio
 - Vector sums
- New methodologies
 - Topic Modelling / Term
 - Topic Modelling / Corpus
- Complete the sentiment analysis...

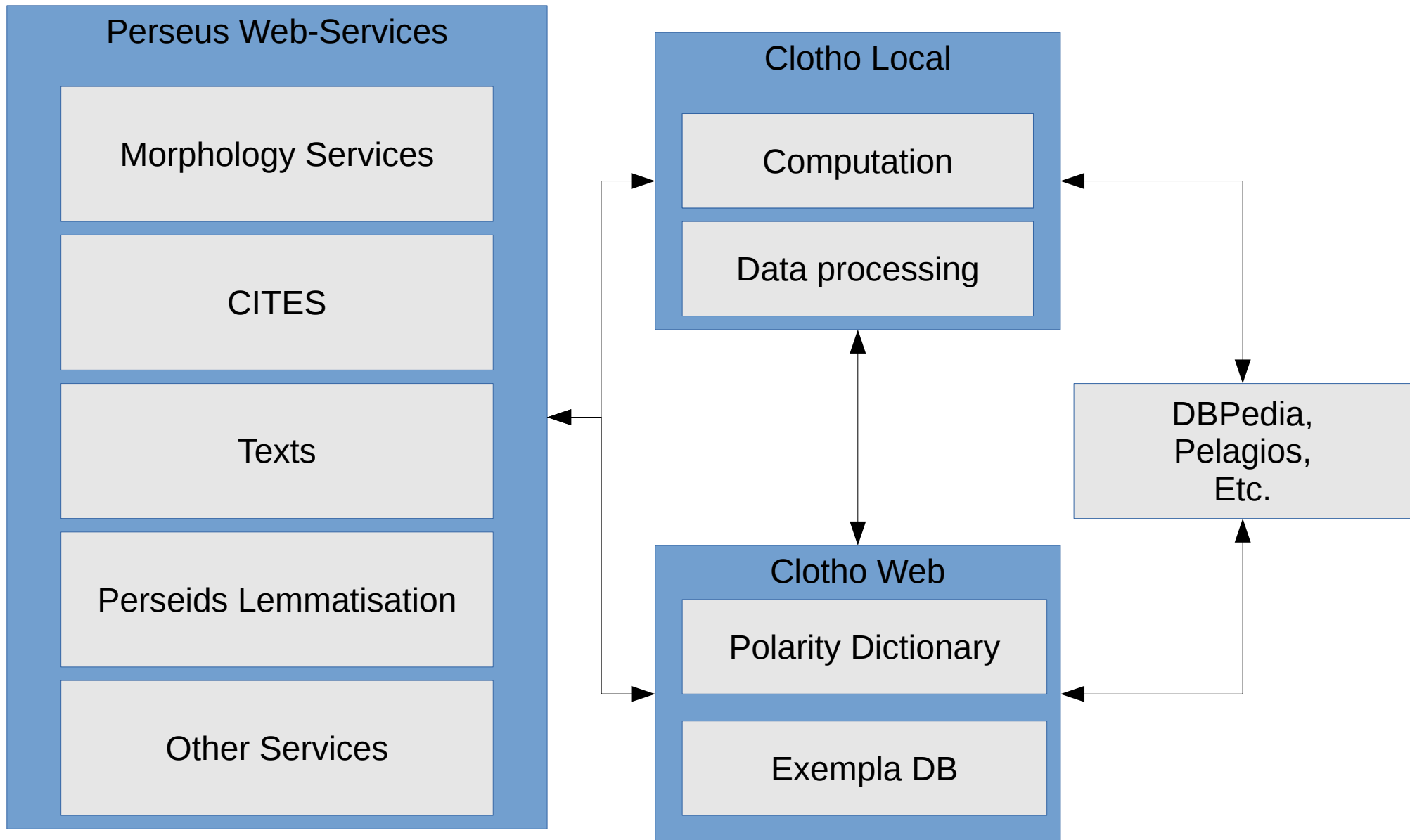
Future : Sustainability

- Code cleaning
- Data
- Instructions and docs
- Python 3 ?

Future : Collaboration

- CLTK ?
- Perseus ?
- Algo addition through discussion with M. Munson ?

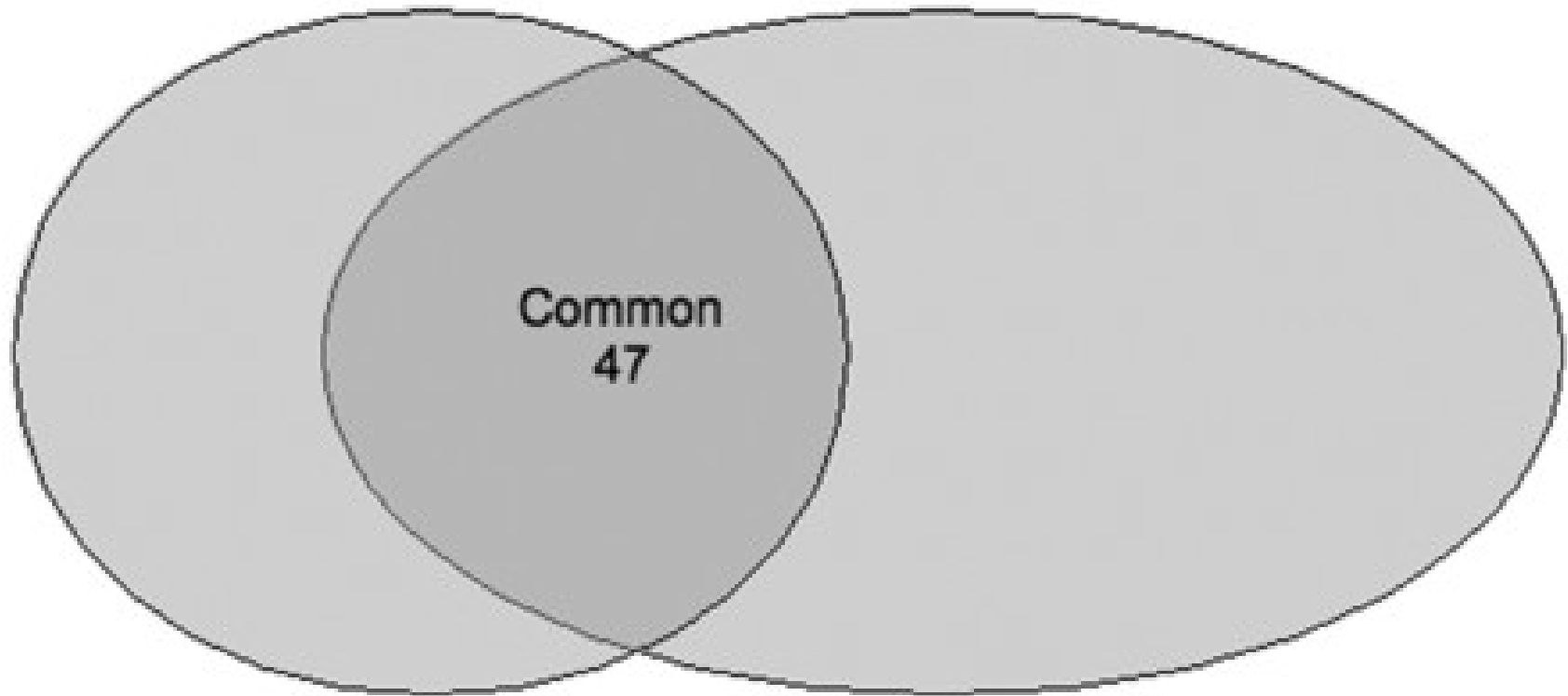
Future : Sustainability AND collaboration



Futur : *Benchmarking ?*

Tesserae
93 Type 4–5

Commentaries
172 Type 4–5



Conclusion

- Connecting the existing tools, providing a pipeline to everyone
- More open corpus = better results
- Those methodologies do not provide answer but they do something a human brain is not able to
- With better annotated (including subject recall for verbs and pronouns), better network
 - The machine can't do that...

Thanks !

- <http://lascivaroma.com>
- <http://github.com/ponteineptique/clotho-local>
- <http://github.com/ponteineptique/clotho-web>
- @ponteineptique
- Thibault.clerice@kcl.ac.uk