Textometric Exploitation of Coreference-annotated Corpora with TXM

Methodological Choices and First Outcomes

Matthieu Quignard (CNRS/ICAR, Lyon)
Serge Heiden (ENS/IHRIM, Lyon)
Frédéric Landragin (CNRS/LATTICE, Paris)
Matthieu Decorde (ENS/IHRIM, Lyon)

Co-reference ?

We had dinner yesterday evening with Serge and Pascal. They were very joyful.

> [They] refers to two people, Serge and Pascal

> [They] and [Serge and Pascal] corefer.

> [They] → [Serge and Pascal] is an anaphora

The wine was delicious.

> [The wine] does not stricty refer to the dinner but to the wine served at dinner (implicit)

> [The wine] → [the dinner] is an associative anaphora

The DEMOCRAT Project

- 3 French partners

  LATTICE (Paris), LILPA (Strasbourg), IHRIM-ICAR (Lyon)
- 48 months

- A manually annotated corpus of French written texts, from 9th century to the 21st ; 1 million words (actually less) ; balanced between narrative and non narrative texts (essays…).
- Tools for annotating texts and exploiting annotation (TXM)
- Support research on reference in discourse and discourse processing
  - Theory of reference (Landragin, Schnedecker)
  - Automated discourse analysis, NLP, deep learning

Annotation principles

- Units – Relations – Schemata (URS, cf. Glozz)
- Units
  - segments of text referring to a given character, idea, concept...
- Relations
  - Units in relation which each other, e.g. anaphora
- Schema
  - Sets of units coreffering to the same object = coreferring chains

Annotation principles

- Units – Relations – Schemata (URS)

- Units
  - segments of text referring to a given character, idea, concept...

- Relations
  - Units in relation which each other, e.g. anaphora

- Schema
  - Sets of units coreffering to the same object = coreferring chains

The « Unitizing » Task

- The annotator must do the following operations
  - Decide whether an object has been mentionned or not
    - Some pronouns are not referential (one, nobody…)
  - **Delimitate** the segment of text that mentions that object
    - The segment has to be contiguous
    - Long enough (whole noun phrase)
    - The preposition should keep outside
  - Give a name to the referent
    and use the **same name** in all coreferring mentions
- NB : There are **overlapping** segments : $[[my]_i$ computer$]_j$

TXM as an annotation framework

http://textometrie.org

- TXM is well known for exploiting, investigating corpora
  - frequencies
  - concordancer
  - charts with R
  - progressions…

- An extension has been developped for DEMOCRAT : ANALEC
  - **Online annotation**
  - Integration of the **URS** structure of annotation over the XML structure of the document itself
  - Online exploitation of the annotation as a means of **verification tool** (quality measurement)

Annotation

et sa felicité dans le paradis, dont la memoire s'est conservée dans l'age d'or des poëtes ; le précepte divin donné à nos premiers parens ; la malice de l'esprit tentateur, et son apparition sous la forme du serpent ; la chute d'**Adam** et d'Eve, funeste à toute leur posterité ; le premier homme justement puni dans tous ses enfans, et le genre humain maudit de Dieu ; la premiere promesse de la rédemption, et la victoire future des hommes sur le démon qui les a perdus.

La terre commence à se remplir, et les crimes s'augmentent. Caïn le premier enfant d'Adam et d'Eve, fait voir au monde naissant la premiere action tragique ; et la vertu commence deslors à estre persecutée par le vice. Là paroissent les moeurs contraires des deux freres : l'innocence d'Abel, sa vie pastorale, et ses offrandes agréables ; celles de Caïn rejettées, son avarice, son impieté, son parricide, et la jalousie mere des meurtres : le

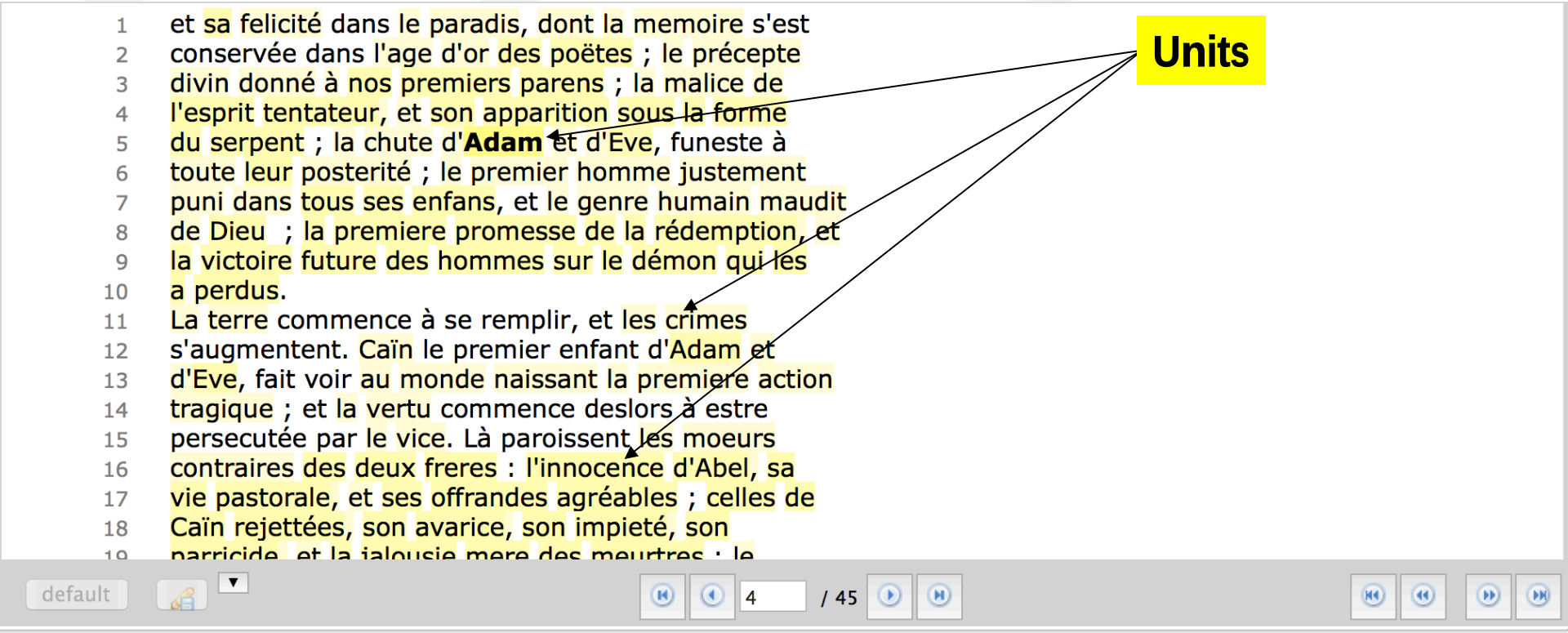
default

4 / 45

DEMOCRATLYON3 - MENTION [443] Console

ACCESSIBILIT Faible CATEGORIE GN.NAM REF Adam

Annotation


et sa felicité dans le paradis, dont la memoire s'est
conservée dans l'age d'or des poëtes ; le précepte
divin donné à nos premiers parens ; la malice de
l'esprit tentateur, et son apparition sous la forme
du serpent ; la chute d'Adam et d'Eve, funeste à
toute leur posterité ; le premier homme justement
puni dans tous ses enfans, et le genre humain maudit
de Dieu ; la premiere promesse de la rédemption, et
la victoire future des hommes sur le démon qui les
a perdus.
La terre commence à se remplir, et les crimes
s'augmentent. Caïn le premier enfant d'Adam et
d'Eve, fait voir au monde naissant la premiere action
tragique ; et la vertu commence deslors à estre
persecutée par le vice. Là paroissent les moeurs
contraires des deux freres : l'innocence d'Abel, sa
vie pastorale, et ses offrandes agréables ; celles de
Caïn rejettées, son avarice, son impieté, son
Units
default
4
/ 45
DEMOCRATLYON3 - MENTION [443]
Console
ACCESSIBILIT
Faible
CATEGORIE
GN.NAM
REF
Adam

Annotation

et sa felicité dans le paradis, dont la memoire s'est conservée dans l'age d'or des poëtes ; le précepte divin donné à nos premiers parens ; la malice de l'esprit tentateur, et son apparition sous la forme du serpent ; la chute d'**Adam** et d'Eve, funeste à toute leur posterité ; le premier homme justement puni dans tous ses enfans, et le genre humain maudit de Dieu ; la premiere promesse de la rédemption, et la victoire future des hommes sur le démon qui les a perdus.

La terre commence à se remplir, et les crimes s'augmentent. Caïn le premier enfant d'Adam et d'Eve, fait voir au monde naissant la premiere action tragique ; et la vertu commence deslors à estre persecutée par le vice. Là paroissent les moeurs contraires des deux freres : l'innocence d'Abel, sa vie pastorale, et ses offrandes agréables ; celles de Caïn rejettées, son avarice, son impieté, son parricide, et la jalousie mere des meurtres ; le

default 4 / 45

DEMOCRATLYON3 - MENTION [443] Console

Unit properties

ACCESSIBILIT Faible CATEGORIE GN.NAM REF Adam

Concordancer

Calling all units coreferring to the same referent

= a view of a referring chain as a sequence of units

text_id	Contexte gauche	Pivot	Contexte droit
MontesquieuLois	, est, à certains égards,	le monarque	; à certains autres, il est le sujet.
MontesquieuLois	le sujet. Il ne peut être	monarque	que par ses suffrages qui sont ses volontés. L
MontesquieuLois	dans une monarchie de sçavoir quel est	le monarque	, et de quelle manière il doit gouverner. Liban
MontesquieuLois	le monarque, et de quelle manière	il	doit gouverner. Libanius dit que à Athènes un
MontesquieuLois	mieux dans la place publique, qu'	un monarque	dans son palais. Mais sçaura – t – il
MontesquieuLois	place publique, qu'un monarque dans	son	palais. Mais sçaura – t – il conduire une
MontesquieuLois	, les sujets sont à l'égard	du monarque	. On n'y doit point donner le suffrage par
MontesquieuLois	accommodées ; le principe du gouvernement arrête	le monarque	; mais, dans une république où un citoyen se
MontesquieuLois	– à – dire de celui où	un seul	gouverne par des loix fondamentales. (j'ai dit
MontesquieuLois	en effet, dans la monarchie,	le prince	est la source de tout pouvoir politique et civil.
MontesquieuLois	dans la monarchie, le prince est	la source de tout pouvoir politique et civil	. Ces loix fondamentales supposent nécessair
MontesquieuLois	que la volonté momentanée et capricieuse d'	un seul	, rien ne peut être fixe, et par conséquent
MontesquieuLois	un bon sujet de défendre la justice	du prince,	ou les limites qu'elle s'est de tout temps
MontesquieuLois	où elles seroient ensevelies. Le conseil	du prince	n'est pas un dépôt convenable. Il est,
MontesquieuLois	, le dépôt de la volonté momentanée	du prince	qui exécute, et non pas le dépôt des loix
MontesquieuLois	dépôt de la volonté momentanée du prince	qui	exécute, et non pas le dépôt des loix fondame
MontesquieuLois	fondamentales. De plus, le conseil	du monarque	change sans cesse ; il n'est point permanent ;
MontesquieuLois	brigues pour être le premier esclave ;	le prince	seroit obligé de rentrer dans l'administration.
MontesquieuLois	aura d'abord la même puissance que	lui	. L'établissement d'un vizir est, dans cet
MontesquieuLois	, et plus, par conséquent,	le prince	est enivré de plaisirs. Ainsi, dans ces états
MontesquieuLois	Ainsi, dans ces états, plus	le prince	a de peuples à gouverner, moins il pense au

Concordancer

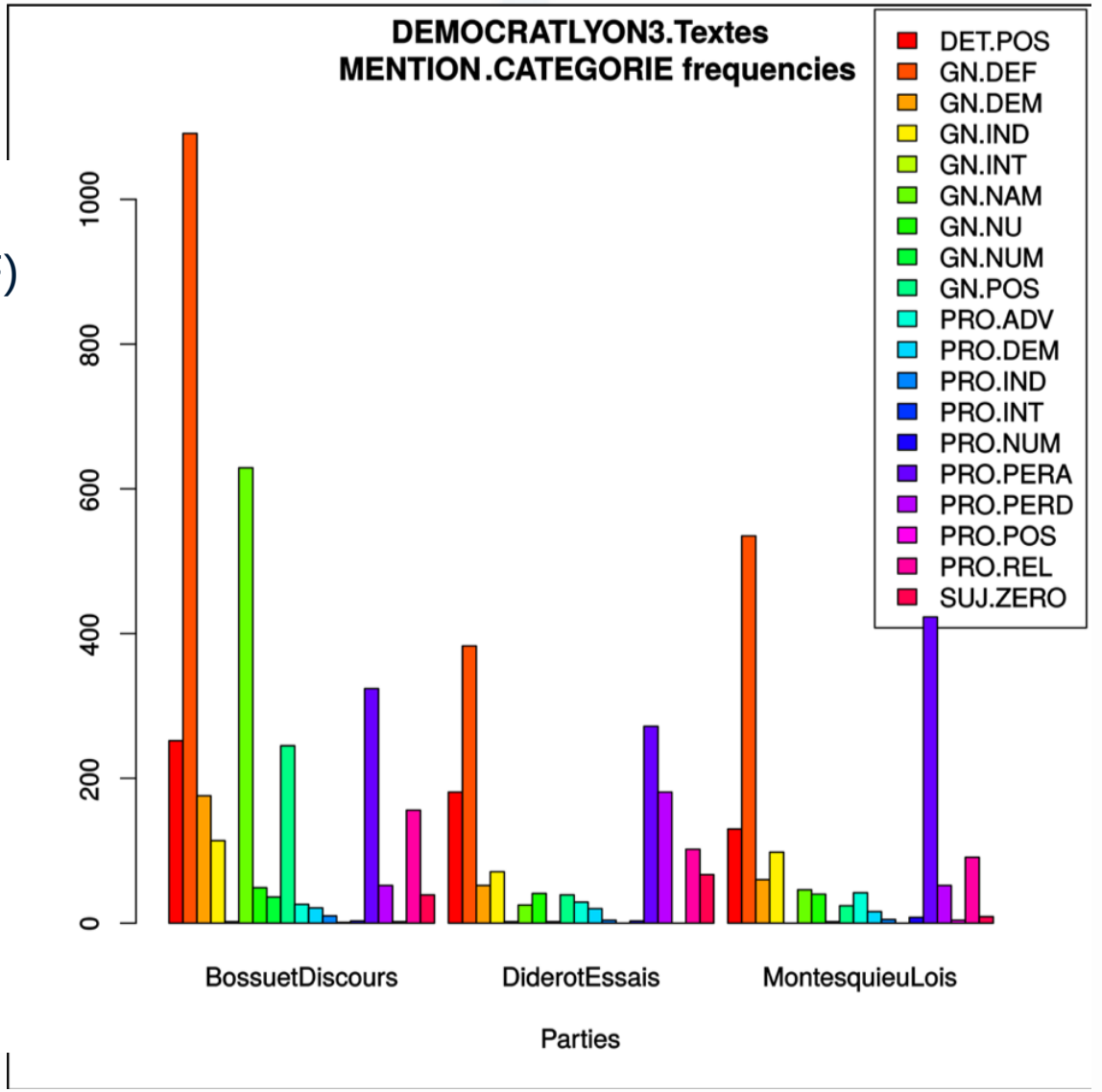
Calling all units coreferring to the same referent

= a view of a referring chain as a sequence of units

text_id	Contexte gauche	Pivot	Contexte droit
MontesquieuLois	, est, à certains égards,	le monarque	; à certains autres, il est le sujet.
MontesquieuLois	le sujet. Il ne peut être	monarque	que par ses suffrages qui sont ses volontés. L
MontesquieuLois	dans une monarchie de sçavoir quel est	le monarque	, et de quelle manière il doit gouverner. Liban
MontesquieuLois	le monarque, et de quelle manière	il	doit gouverner. Libanius dit que à Athènes un
MontesquieuLois	mieux dans la place publique, qu'	un monarque	dans son palais. Mais sçaura – t – il
MontesquieuLois	place publique, qu'un monarque dans	son	palais. Mais sçaura – t – il conduire une
MontesquieuLois	, les sujets sont à l'égard	du monarque	. On n'y doit point donner le suffrage par
MontesquieuLois	accommodées ; le principe du gouvernement arrête	le monarque	; mais, dans une république où un citoyen se
MontesquieuLois	– à – dire de celui où	un seul	gouverne par des loix fondamentales. (j'ai dit
MontesquieuLois	en effet, dans la monarchie,	le prince	est la source de tout pouvoir politique et civil.
MontesquieuLois	dans la monarchie, le prince est	la source de tout pouvoir politique et civil	. Ces loix fondamentales supposent nécessair
MontesquieuLois	que la volonté momentanée et capricieuse d'	un seul	, rien ne peut être fixe, et par conséquent
MontesquieuLois	un bon sujet de défendre la justice	du prince,	ou les limites qu'elle s'est de tout temps
MontesquieuLois	où elles seroient ensevelies. Le conseil	du prince	n'est pas un dépôt convenable. Il est,
MontesquieuLois	, le dépôt de la volonté momentanée	du prince	qui exécute, et non pas le dépôt des loix
MontesquieuLois	dépôt de la volonté momentanée du prince	qui	exécute, et non pas le dépôt des loix fondame
MontesquieuLois	fondamentales. De plus, le conseil	du monarque	change sans cesse ; il n'est point permanent ;
MontesquieuLois	brigues pour être le premier esclave ;	le prince	seroit obligé de rentrer dans l'administration.
MontesquieuLois	aura d'abord la même puissance que	lui	. L'établissement d'un vizir est, dans cet
MontesquieuLois	, et plus, par conséquent,	le prince	est enivré de plaisirs. Ainsi, dans ces états
MontesquieuLois	Ainsi, dans ces états, plus	le prince	a de peuples à gouverner, moins il pense au

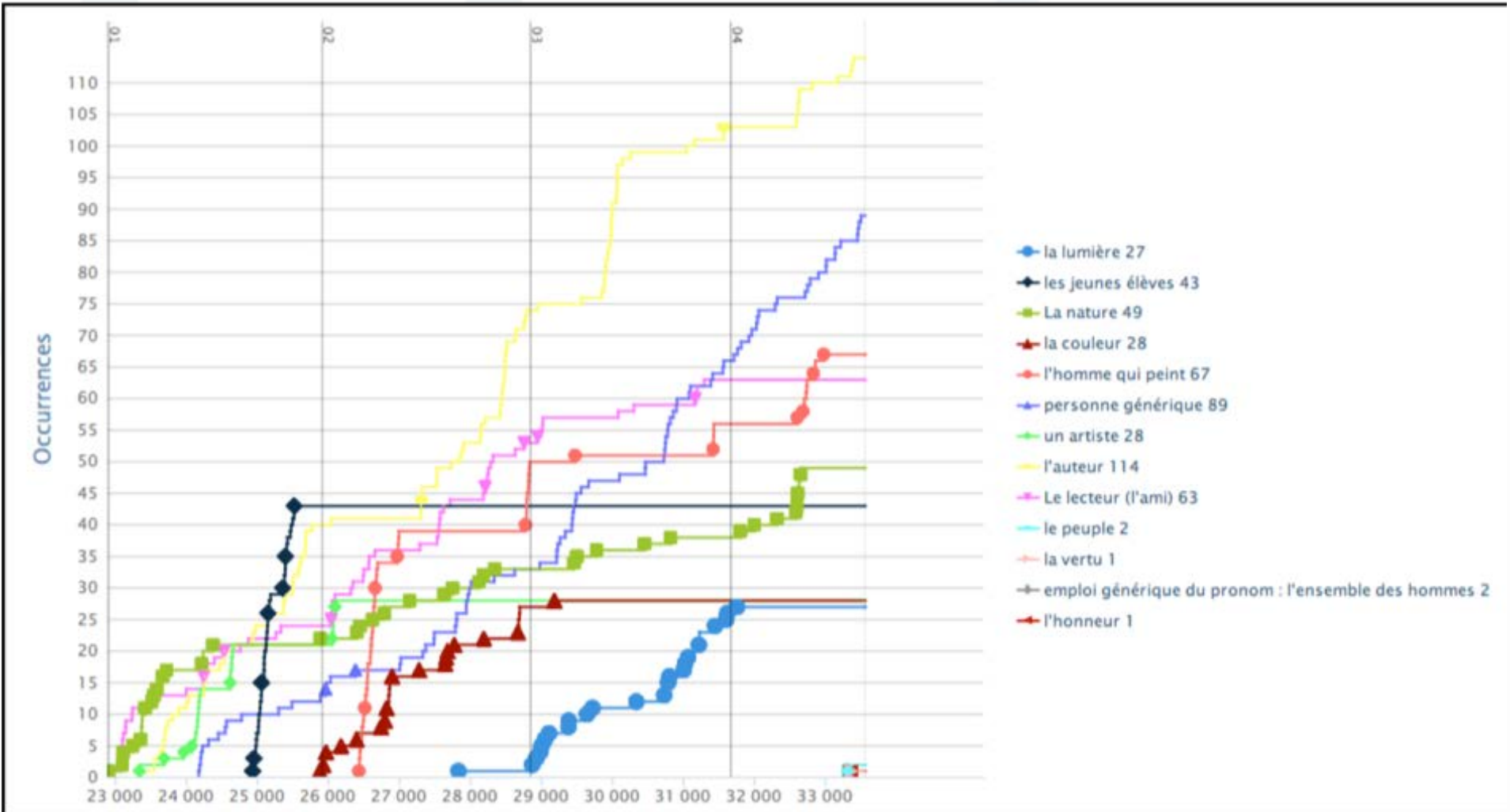
Histogram

- Common pattern
  - Definite NP• (GN.DEF)
  - Personal Pronouns •• (PRO.PERx)
  - Relative Pronouns (PRO.REL) •
- Pattern specific of Bossuet's Discourse
  - Lots of Proper Nouns (GN.NAM) •
  - NP with possessives (GN.POS) •



Progressions

= view of chains (schemata) throughout the corpus across chapters (textual units)



Summary

- URS annotation is now possible in TXM => complex annotation tasks such as semantic, reference tagging
- URS annotation can also be queried and checked via usual TXM tools
  - direct feedback on what we are currently doing
  - quality checking, consistancy
- Undergoing work
  - develop groovy scripts for checking errors and inconsistancies
  - embed algorithms for intercoder reliability
    - at unit level (segmentation)
    - at schema level (referent identification)