Textometric Exploitation of Coreference-annotated Corpora with TXM: Methodological Choices and First Outcomes

Matthieu Quignard[1], Serge Heiden[2], Frédéric Landragin[3], Matthieu Decorde[2]

[1]ICAR, CNRS, University of Lyon – matthieu.quignard@ens-lyon.fr

[2]IHRIM, ENS Lyon, CNRS, University of Lyon – {slh,matthieu.decorde}@ens-lyon.fr

[3]Lattice, CNRS, ENS Paris, University Sorbonne Nouvelle, PSL Research University, USPC – frederic.landragin@ens.fr

AUTHORS' DRAFT

Abstract

In this article we present a set of measures – some of which can lead to specific visualisations – with the objective to enrich the possibilities of exploration and exploitation of annotated data, and in particular coreference chains. We first present a specific use of the well-known concordancer, which is here adapted to present the elements of a coreference chain. We then present a histogram generator that allows for example to display the distribution of the various coreference chains of a text, given a value from the annotated properties. Finally, we present what we call progress diagrams, whose purpose is to display the progress of each chain throughout the text. We conclude on the interest of these (interactive) modes of visualization in order to make the annotation phase more controlled and more effective.

Résumé

Nous présentons dans cet article un ensemble de mesures – dont certaines peuvent amener à des visualisations spécifiques – dont l'objectif est d'enrichir les possibilités d'exploration et d'exploitation des données annotées, en particulier quand il s'agit de chaînes de coréférences. Nous présentons tout d'abord une utilisation adaptée de l'outil bien connu qu'est le concordancier, en n'affichant que les maillons d'une chaîne choisie. Puis nous montrons un générateur d'histogramme qui permet par exemple d'afficher la répartition des chaînes de coréférences d'un texte à partir d'une propriété annotée. Nous montrons enfin ce que nous appelons des diagrammes de progression, dont le but est d'afficher les avancées au fur et à mesure du texte des chaînes de coréférences qu'il contient. Nous concluons sur l'intérêt de ces modes (interactifs) de visualisation pour rendre la phase d'annotation plus maîtrisée et plus efficace.

Keywords: coreference chain, corpus annotation, annotation tool, visualisation tool, exploration tool, statistical analysis of textual data.

1. Introduction

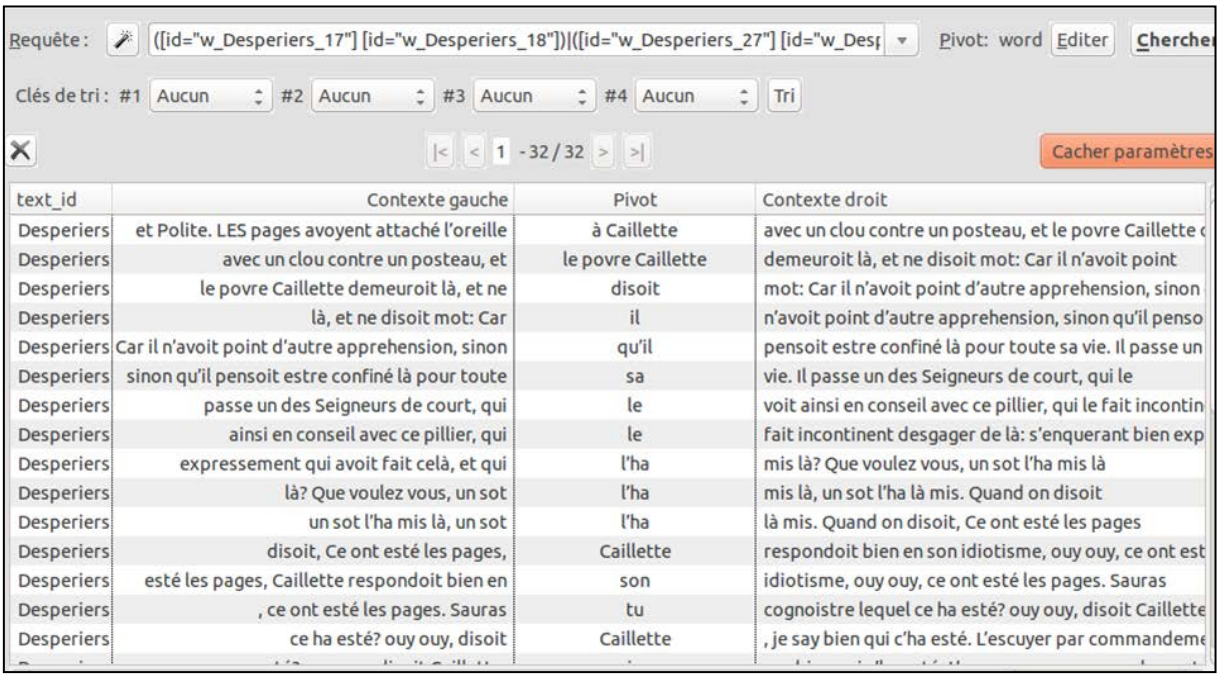
The manual annotation of a textual corpus with referring expressions (Charolles, 2002) and coreference chains (Schnedecker, 1997, Landragin & Schnedecker, 2014) requires adapted tools. A coreference chain can cover the whole text; it is therefore a linguistic object for which the existing means of visualization and exploration are few and often perfectible. The MMAX2 tool (Müller & Strube, 2006) allows for visualizing the links between

referring expressions using arrows which link markables. The GLOZZ tool (Mathet & Wildlöcher, 2009) offers several means of visualization: with arrows like MMAX2, or with a specific marking in the margin or the middle of the text. The ANALEC tool (Landragin *et al.*, 2012) and its specific extension for coreference chains (Landragin, 2016) proposes a graphic metaphor based on the succession of coloured dots. This allows the analyst to configure visual parameters, for instance the colour which can be linked to any of the annotated properties. This type of visualization makes it possible to see at a glance the structural differences between the different reference chains of a text. That must be useful to the analyst, in addition to manual explorations and finer linguistic analyses.

2. Linguistic objects and methodology

In the continuity of previous works (Heiden, 2010; Landragin, 2016), we present here a set of measures – some of which can lead to specific visualisations – with the objective to enrich the possibilities of exploration and exploitation of annotated data. We focus in particular on annotations which concern discursive phenomena like coreference, i.e., annotations which are necessarily described within two levels: 1. markable, group of contiguous words to which is assigned some labels, using for instance a feature structure; 2. set of markables, or links between markables, as is it the case for any chain of annotations: anaphoric chains, textual organizers chains, textual structure elements chains, etc. A feature structure can also be assigned at level 2, i.e., to the set or to the links.

3. A concordancer adapted to annotations chains



text_id	Contexte gauche	Pivot	Contexte droit
Desperiers	et Polite. LES pages avoyent attaché l'oreille	à Caillette	avec un clou contre un posteau, et le povre Caillette (
Desperiers	avec un clou contre un posteau, et	le povre Caillette	demeuroit là, et ne disoit mot: Car il n'avoit point
Desperiers	le povre Caillette demeuroit là, et ne	disoit	mot: Car il n'avoit point d'autre apprehension, sinon
Desperiers	là, et ne disoit mot: Car	il	n'avoit point d'autre apprehension, sinon qu'il penso
Desperiers	Car il n'avoit point d'autre apprehension, sinon	qu'il	pensoit estre confiné là pour toute sa vie. Il passe un
Desperiers	sinon qu'il pensoit estre confiné là pour toute	sa	vie. Il passe un des Seigneurs de court, qui le
Desperiers	passe un des Seigneurs de court, qui	le	voit ainsi en conseil avec ce pillier, qui le fait incontin
Desperiers	ainsi en conseil avec ce pillier, qui	le	fait incontinent desgager de là: s'enquerant bien exp
Desperiers	expressement qui avoit fait celà, et qui	l'ha	mis là? Que voulez vous, un sot l'ha mis là
Desperiers	là? Que voulez vous, un sot	l'ha	mis là, un sot l'ha là mis. Quand on disoit
Desperiers	un sot l'ha mis là, un sot	l'ha	là mis. Quand on disoit, Ce ont esté les pages
Desperiers	disoit, Ce ont esté les pages,	Caillette	respondoit bien en son idiotisme, ouy ouy, ce ont est
Desperiers	esté les pages, Caillette respondoit bien en	son	idiotisme, ouy ouy, ce ont esté les pages. Sauras
Desperiers	, ce ont esté les pages. Sauras	tu	cognoistre lequel ce ha esté? ouy ouy, disoit Caillette
Desperiers	ce ha esté? ouy ouy, disoit	Caillette	, je say bien qui c'ha esté. L'escuyer par commandeme

Fig 1: Concordancer with the elements of a coreference chain, dedicated to a character named "Caillette".

As a first visualization mode, we reuse the very classic concordancer to display the elements which constitute a coreference chain. The use of such a visualization tool, which is well established in the community of corpus exploration (Poudat & Landragin, 2017), seemed natural for visualizing chains of annotations. The last version of TXM (Heiden, 2010) thus includes a concordancer which makes it possible to display in a column all the

elements (e.g. referring expressions) of a chain (e.g. coreference chain), with left and right contexts for each elements. Compared to MMAX2 (Müller & Strube, 2006) and GLOZZ (Mathet & Wildlöcher, 2009) visualisation choices, i.e. arrows linking marquables which are displayed directly on the text, this concordancer has the advantage of regrouping all the relevant information in a small graphic space.

Fig. 1 shows the list of all referring expression to the character 'Caillette'. Sorted in the textual order, the concordancer shows the alternation of the use of proper nouns, pronouns, possessives, etc. This concordancer may also be sorted along a given property of the marquable, e.g. its POS label. This representation may then be exploited to see whether the POS annotation is consistent or not.

4. Histograms for visualising distributions of annotations chains

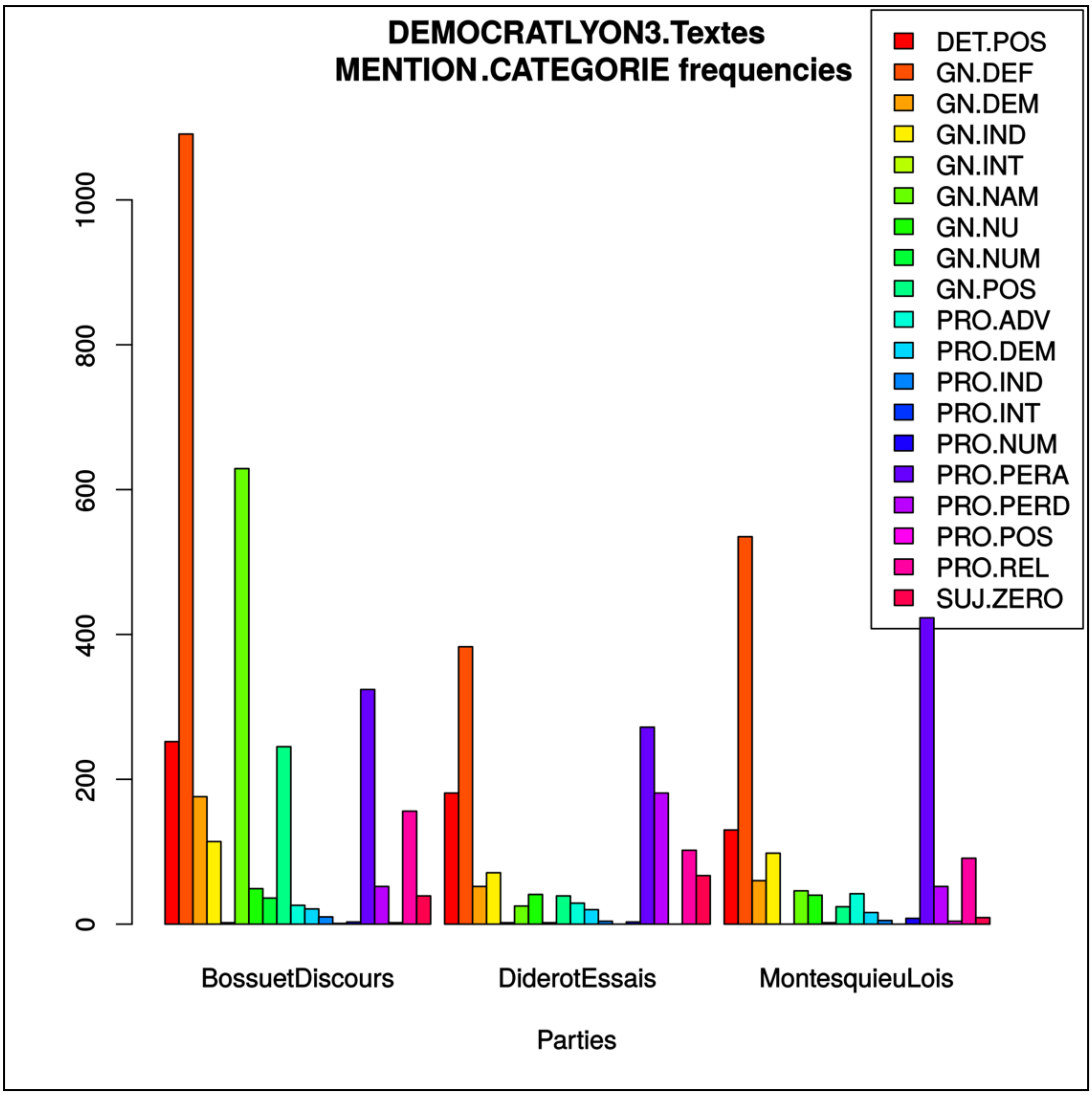


Fig 2: Comparative barplots of grammatical categories usage by reference units in three texts: Bossuet, "Discours sur l'histoire universelle" (1681), Diderot, "Essais sur la peinture" (1759-1766), Montesquieu, "Esprit des lois" (1755).

A second mode of visualization, also very traditional, is the histogram (bar plot). The user can select one or several properties – the determination of the referring expressions, for

instance, or the type of referent – and launch calculations on their occurrences: cross-counts, correlation computation and so on. TXM now includes a histogram generator, which allows for example to display the distribution of coreference chains throughout the text, as well as the distribution of chains according to the number of referring expressions they include. These calculations and their associated visualizations provide TXM with integrated functionalities which required in other state-of-the-art tools the development of scripts, in order to export the relevant data and exploit them in an external tool like a spreadsheet.

Figure 2 compares the distribution of grammatical categories of referring expressions in three texts. Although all texts are all encyclopedical ones, the Discourse from Bossuet shows a particular profile, with a high number of proper nouns (GN.NAM).

5. Progression charts for annotations chains

A third (new) mode of visualization consists to graphically show the progress of each chain throughout the text. The principle is simple, but the possibilities of exploration and exploitation of the generated graph are numerous. In a two-dimensional chart the abscissa of which represents the linearity of the text, chains are displayed point by point (cf. Fig. 3): each occurrence of a referring expression increases by one notch the ordinate of the corresponding point. The resulting broken lines are all ascending but can considerably vary in their areas of progression and flat areas.

When they are visualized simultaneously, it is possible to detect the parts of the text where several referents are competitors, or on the contrary those where several referents appear alternately. Zooming (in and out) as well as focussing features allows for visualizing the characteristics of each point, thus enriching the exploration possibilities of these progression chart and the underlying coreference chains.

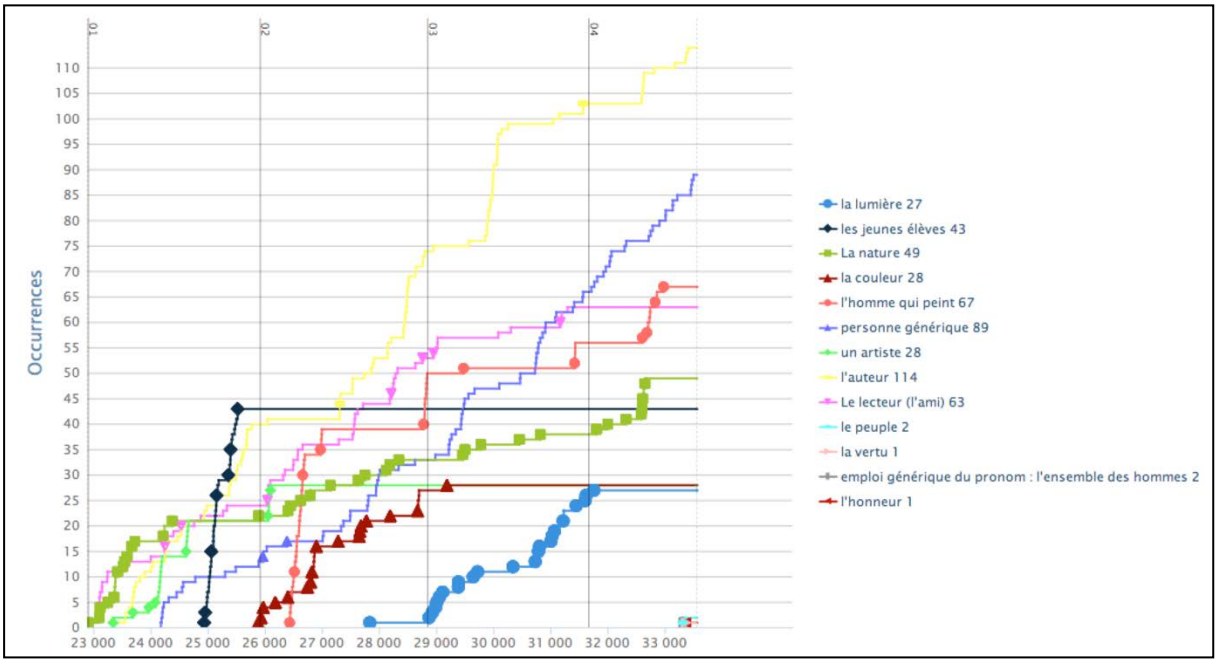


Fig 3: Progression graph of the main coreference chains at the beginning of "Essais sur la peinture" from Denis Diderot. The dots highlighted with symbols correspond to referring expressions with low accessibility.

6. Discussion

The common points of these new visualization modes is not only to propose visual representations which are easy to understand (and possibly interactive, when it is possible to modify on the fly one of the properties), to allow the visualization of these representations directly in TXM, with no need to export annotated data and to use external tools, but also to facilitate the detection by the analyst of intruders, outliers and deviant examples. For instance potential annotation errors: it can be the case for a referring expression which has nothing to do in the currently visualised chain. It may be a peak or a suspect flat in one of the generated histograms. It may be a zone with a very high slope (or a very long flat) in a progression diagram. In all three cases, the analyst can directly access the suspicious annotation, in order to verify it and of course to modify it. The integration of the measurements and their visualizations in TXM allows this immediate return to the corpus annotation phase. This is particularly effective when the corpus is being annotated manually.

7. Conclusion and future works

One can say that it is by annotating that we can see the mistakes we make, but we still need appropriate tools to detect these errors. With the new possibilities of interaction that we propose here, we hope that we are taking a significant step in this direction. The first tests which we have carried out demonstrated the relevance of our approach.

References

Charolles M. (2002). *La référence et les expressions référentielles en français*. Ophrys, Paris, France.

Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, Nov. 2010. Sendai, Japan, Institute for Digital Enhancement of Cognitive Development, Waseda University, pp. 389-398, available at halshs.archives-ouvertes.fr/halshs-00549764.

Landragin F. (2016). Conception d'un outil de visualisation et d'exploration de chaînes de coréférences. *Statistical Analysis of Textual Data – Proceedings of 13th International Conference Journées d'Analyse statistique des Données Textuelles (JADT 2016)*, Nice, France, pp. 109-120.

Landragin F., Poibeau T. and Victorri B. (2012). ANALEC: a New Tool for the Dynamic Annotation of Textual Data. *Proceedings of LREC 2012*, Istanbul, Turkey, pp. 357-362.

Landragin F. and Schnedecker C., editors (2014). *Les chaînes de référence*. Volume 195 of the *Langages* journal, Armand Colin, Paris, France.

Müller C. and Strube M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun S., Kohn K. and Mukherjee J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, Peter Lang, Frankfurt, Germany.

Poudat, C. and Landragin, F. (2017). Explorer un corpus textuel : méthodes, pratiques, outils. Champs Linguistiques. De Boeck Supérieur : Louvain-la-Neuve.

Schnedecker C. (1997). *Nom propre et chaîne de référence*. Klincksieck, Paris, France.

Widlöcher A. and Mathet Y. (2012). The Glozz platform: a corpus annotation and mining tool. In Concolato C. and Schmitz P, editors, *Proceedings of the ACM Symposium on Document Engineering (DocEng'12)*, Paris, France, pp. 171-180.