Un exemple de corpus annoté en diachronie longue :

le corpus Democrat, enjeux et exploitations

Julie Glikman [1]
Frédéric Landragin [2]
Catherine Schnedecker [1]
Amalia Todirascu [1]

[1] Université de Strasbourg, LiLPa, UR1339
[2] CNRS, laboratoire Lattice, UMR8094

ConCorDiaL, 14 octobre 2022

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR

UN CORPUS ANNOTÉ EN CHAINES DE RÉFÉRENCE

Prenons un exemple

Comme tout avait brûlé – la mère, les meubles et les photographies de la mère -, pour Fabre et le fils Paul c'était tout de suite beaucoup d'ouvrage : toute cette cendre et ce deuil, déménager, courir se refaire dans les grandes surfaces. Fabre trouva trop vite quelque chose de moins vaste, deux pièces aux fonctions permutables sous une cheminée de brique dont l'ombre donnait l'heure, et qui avaient ceci de bien d'être assez proches du quai de Valmy.

Le soir après le dîner, Fabre parlait à Paul de sa mère, sa mère à lui Paul, parfois dès le dîner. Comme on ne possédait plus de représentation de Sylvie Fabre, il s'épuisait à vouloir la décrire toujours plus exactement : au milieu de la cuisine naquirent des hologrammes que dégonflait la moindre imprécision. Ça ne se rend pas, soupirait Fabre en posant une main sur sa tête, sur ses yeux, et le découragement l'endormait. Souvent ce fut à Paul de déplier le canapé convertible, transformant les choses en chambre à coucher.

- Une chaîne par référent, sachant qu'un référent peut être un individu (mère, père, fils), mais aussi un groupe d'individus (le père + le fils)
- Une chaîne regroupe des expressions référentielles et (éventuellement) des indices morphologiques rappelant l'existence du référent
- Ambiguïtés référentielles, flou référentiel (pronom « on »), très grande diversité des expressions référentielles, non correspondance entre forme de l'expression et caractère référentiel… Tout ceci rend l'annotation de chaînes de référence complexe, chronophage, ainsi que sujette à désaccords
- Quelle méthodologie mettre en œuvre pour produire un corpus annoté ?
- Comment exploiter ces annotations ?

Annoter les chaînes : pourquoi ?

- Constituer un **corpus de référence** sur la (co)référence, pour le français écrit, en complément du corpus ANCOR (oral transcrit)
- Contribuer ainsi aux **humanités numériques**
- Rendre possible des modes d'interrogation pour permettre des **études linguistiques qualitatives et quantitatives** :
  - comment les chaînes de coréférence sont-elles constituées ?
  - à quel rang dans une chaîne un démonstratif apparaît-il ?
  - quelle est la fréquence d'apparition des noms propres ?
  - quelles sont les expansions des expressions référentielles ?
  - à quoi ressemblent les débuts de chaîne ?
  - peut-on prévoir des motifs dans la manière dont les chaînes se croisent ?
  - y a-t-il corrélation entre un genre textuel et un type de chaîne ?
- Nourrir, par **apprentissage artificiel**, des systèmes de TAL

Un projet pour faire tout ça…

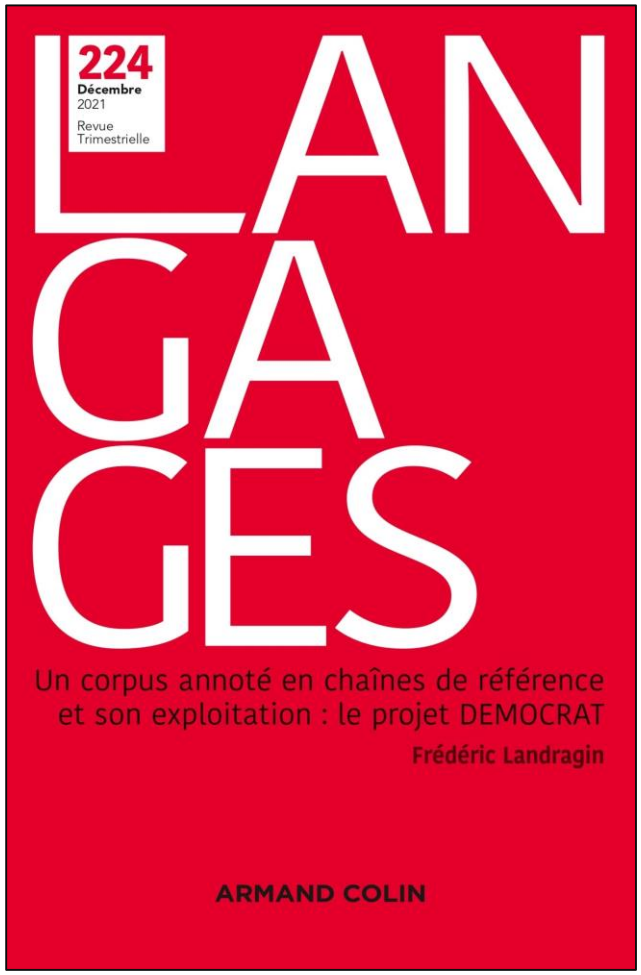
Projet de 4 ans financé par l'ANR (2016-2020)

Site web : http://www.lattice.cnrs.fr/democrat/

3 laboratoires partenaires,
48 participants

ANR-15-CE38-0008

Publications du projet sur le corpus et les études contrastives



https://www.ortolang.fr/market/corpora/democrat/

Un corpus de textes écrits : choix structurels

- 50% français contemporain – 50% autres
- 50% genre narratif – 50% autres
  - Genre narratif : nouvelles, débuts de roman
  - Autres : textes de presse, textes scientifiques, textes de loi, etc.
- Répartition diachronique la plus homogène possible
- Repères quantitatifs :
  - 58 textes ou extraits de textes
  - Chaque texte comporte environ 10.000 mots et reste cohérent (un seul texte, ou un seul auteur)
  - Au total : 689 000 mots
    198 000 expressions référentielles annotées
    20 000 chaînes (dont 9 000 > 2 maillons)

Un corpus de textes écrits : sélection des extraits

- Prérequis indispensable : textes libres de droit
- Réutilisation de textes déjà annotés par ailleurs (par exemple en morphosyntaxe et/ou en syntaxe), en prévision de futures analyses croisées
- Pour les textes en français médiéval, recours à la BFM
- Pour les textes en français contemporain, recours à Gallica
- Phase inévitable de « nettoyage » des textes avant annotation : pas de gestion de couches d'annotation multiples

Le modèle URS de Glozz

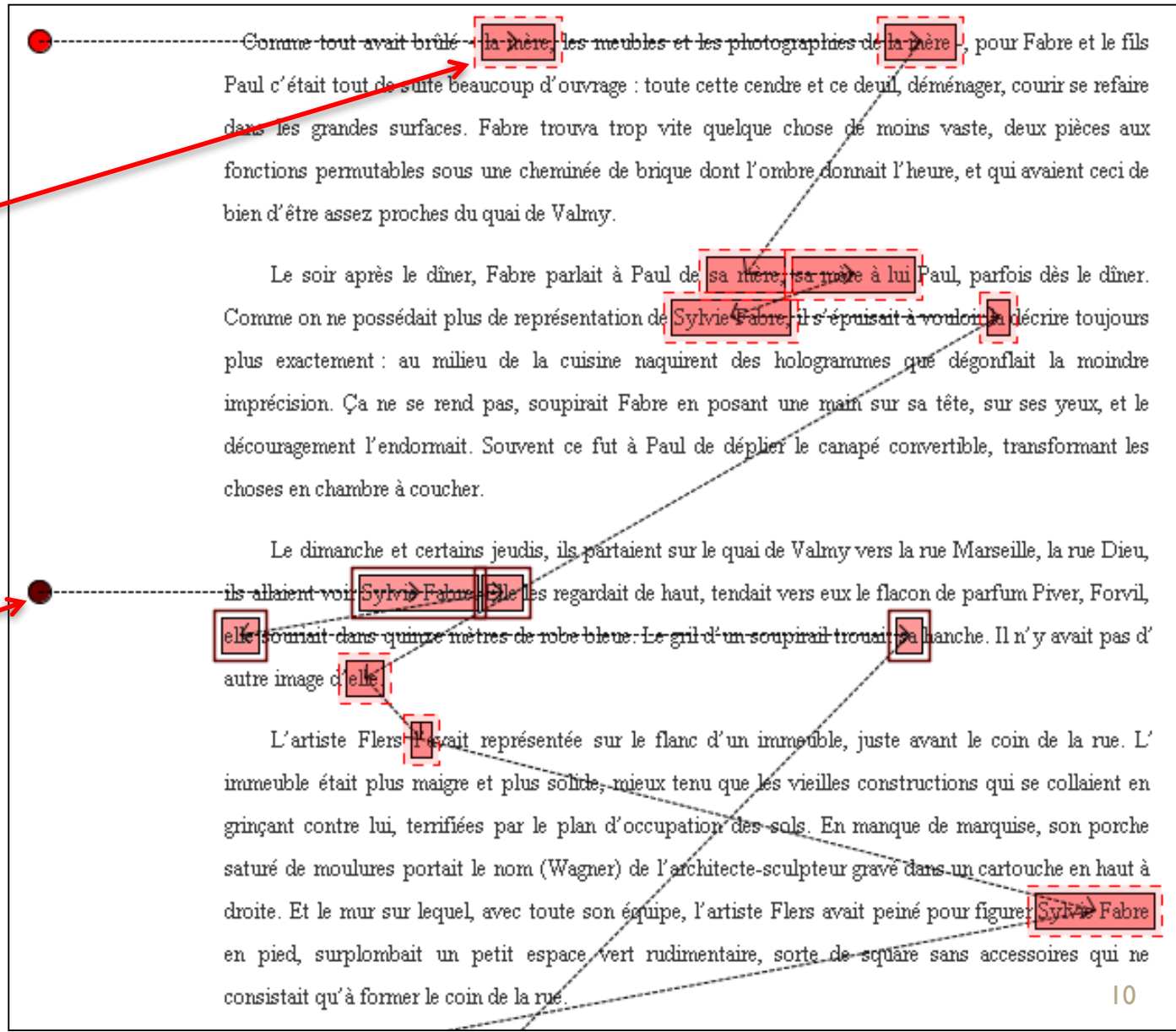
- Glozz, Analec et désormais TXM partagent un même modèle pour la représentation des annotations : URS
  - U = unités : ce sont les marquables
  - R = relations : ce sont des liens orientés entre 2 marquables
  - S = schémas : ce sont des ensembles (hétérogènes) d'unités, de relations et de schémas, qui permettent de modéliser des objets complexes tels que les structures argumentatives ou… les chaînes de coréférences

- Les choix de Democrat
  - Les expressions référentielles font l'objet d'un type d'unité
  - Les chaînes de coréférences font l'objet d'un type de schéma
  - Éventuellement, d'autres objets sont envisageables

Deux annotations différentes

Une unité de type « expression référentielle »

Un schéma de type « chaîne de coréférences »



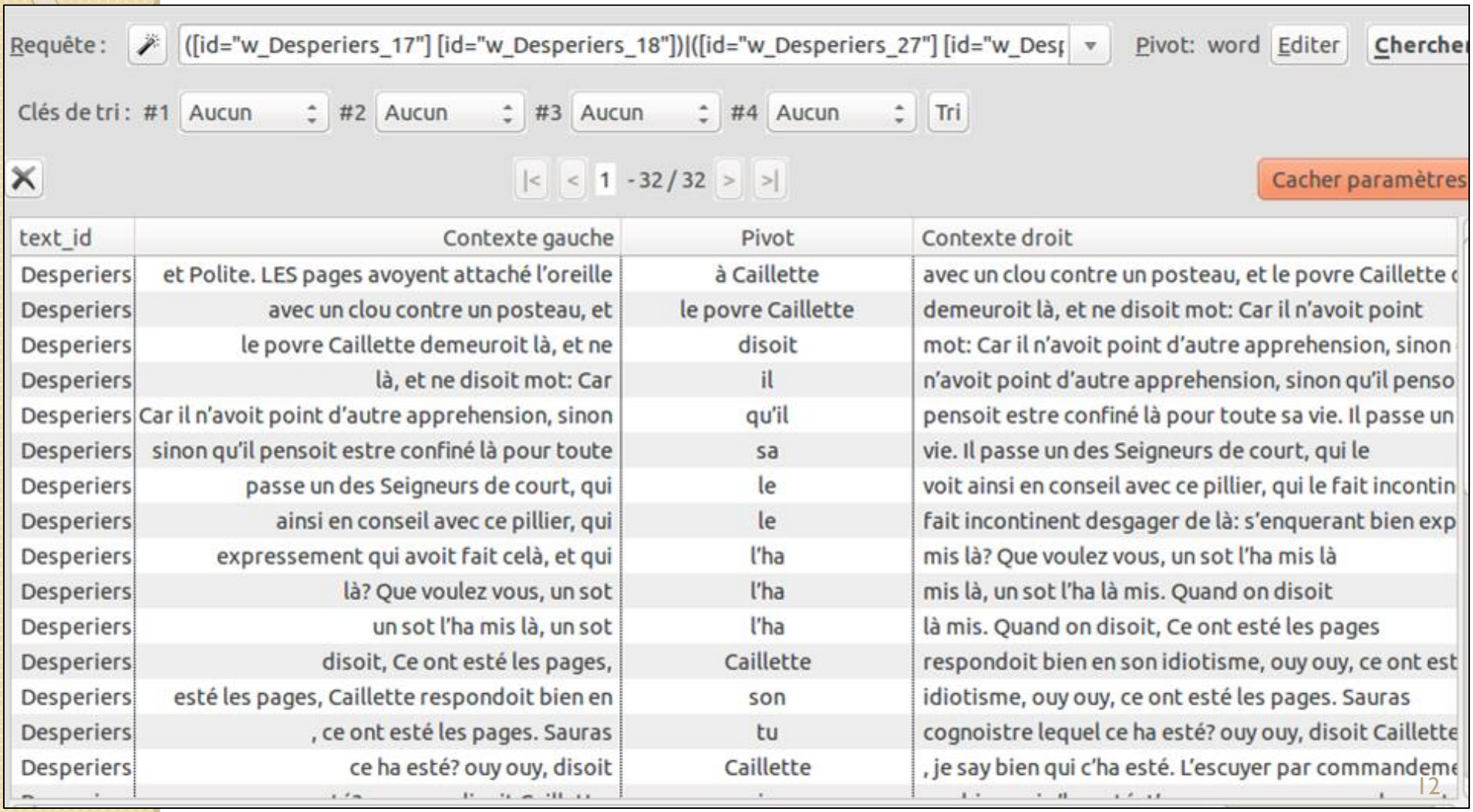
Extension URS de TXM

- But : ajouter à TXM des fonctionnalités d'annotation et de gestion des annotations

- Calculs possibles grâce aux macro TXM
  - Nombre de référents
  - Nombre d'expressions référentielles
  - Densité référentielle (nb réf / nb mots)
  - Nombre de chaînes
  - Longueur des chaînes (en nb de maillons)

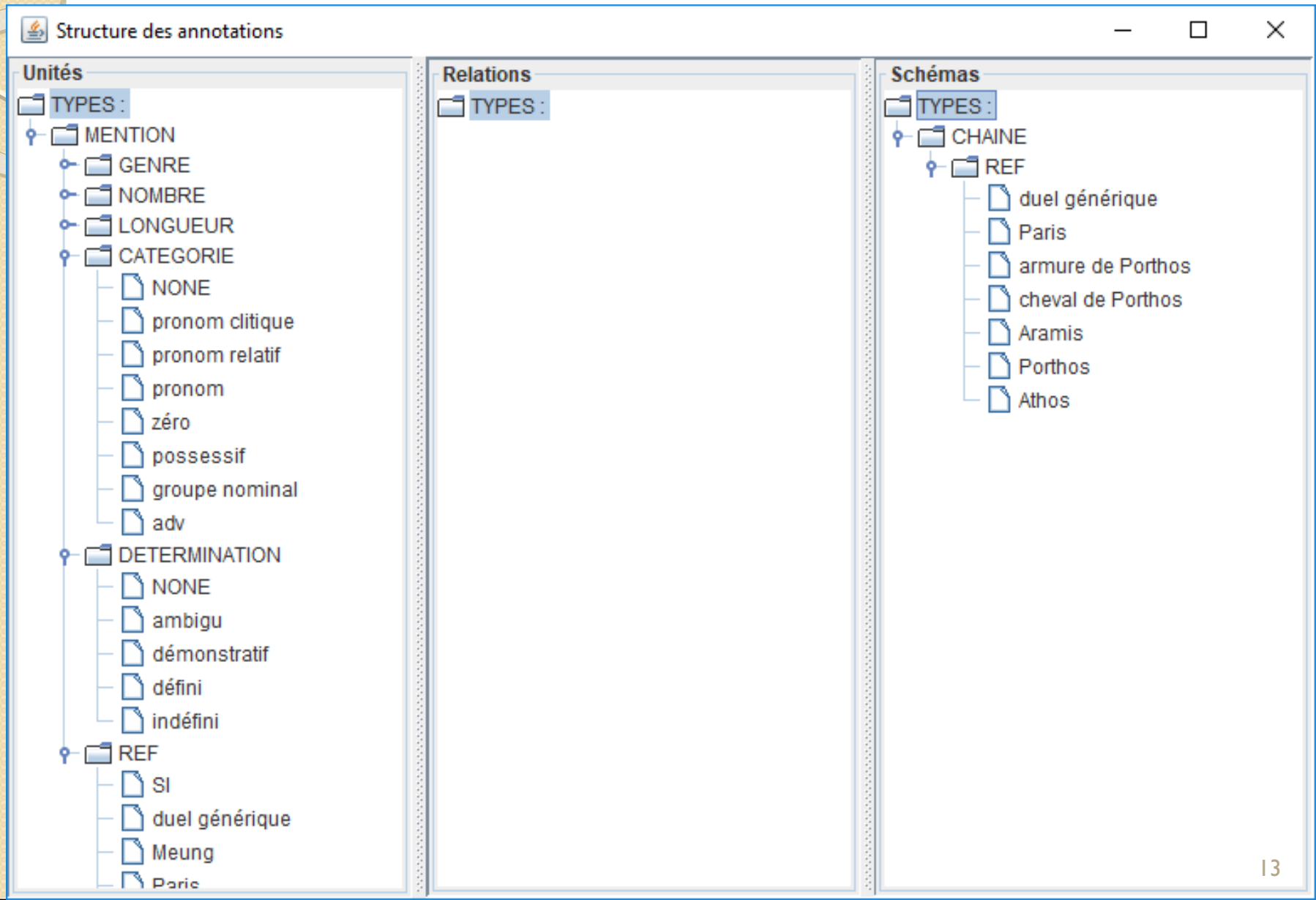
- Annotations supplémentaires automatisables
  - Catégories grammaticales des maillons
  - Détermination

Exemple de nouveauté de TXM : concordancier appliqué aux chaînes



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

Structure URS des annotations



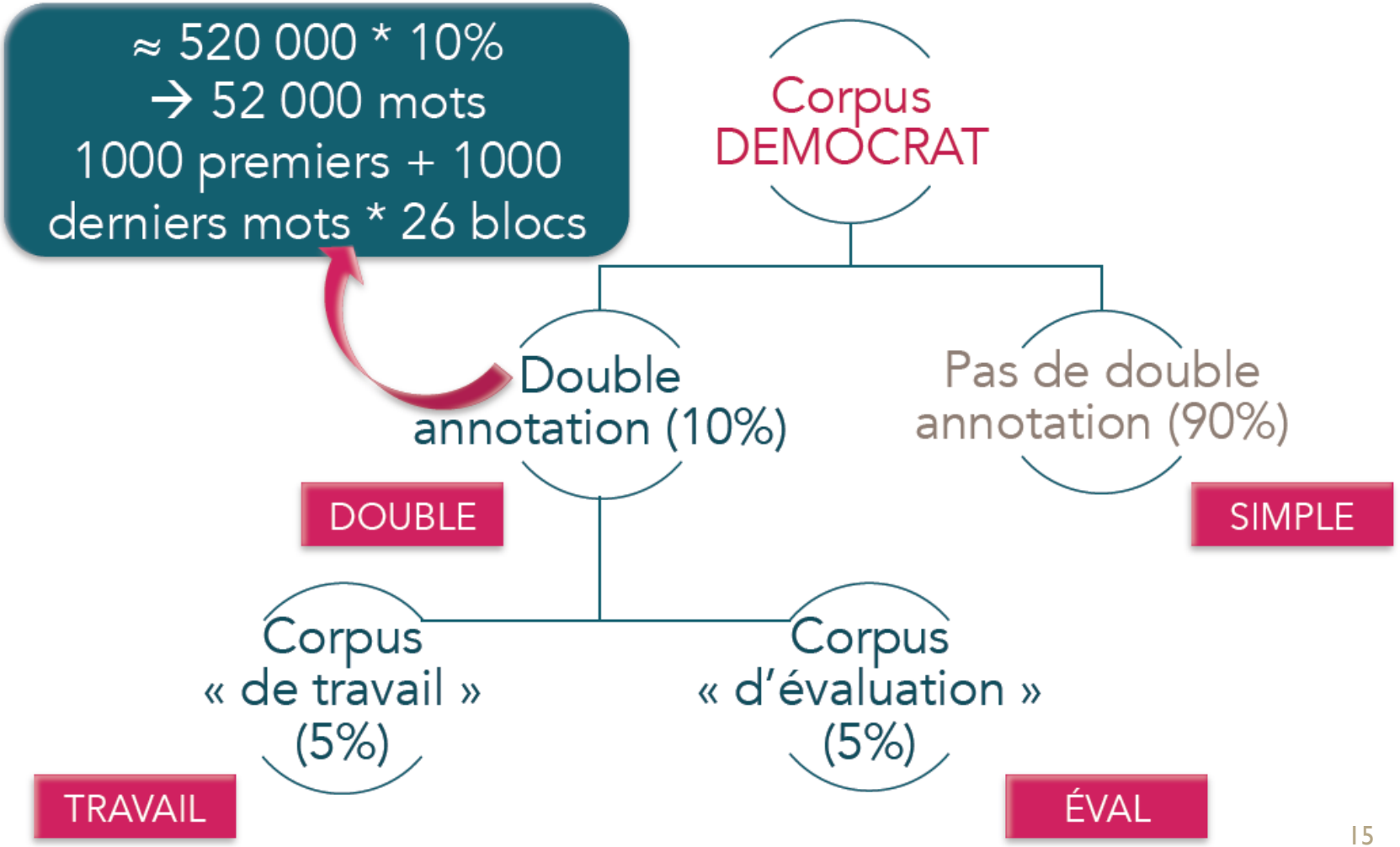
Qu'annote-t-on exactement ?

- Toutes les expressions qui réfèrent, qu'elles désignent des référents humains, animés ou des objets, des notions, des lieux, des dates… à partir du moment où l'expression est de type GN, nom propre ou pronom

- Cas particulier : les événements :
  - [Les vêtements récoltés]$_i$ vont être **vendus** à [la société …]$_j$. [Le produit]$_k$ de **[cette vente]$_l$** sera reversé…
  - Il **neige**. **[Elle]$_i$** tient.

- Cas particulier (important pour l'ancien et le moyen français) Les pronoms zéro : on annote le verbe support :
  - [Pierre]$_a$ boit mais ne [fume]$_a$ pas.
  - [Fermez]$_a$ [la porte]$_b$!

Evaluation de la qualité et scission du corpus



UN EXEMPLE D'EXPLOITATION DIACHRONIQUE

Chaînes de référence et structuration textuelle

- Comment une chaîne commence-t-elle ?
- A quel moment se termine une chaîne ?
  - Doit-on terminer toutes les chaînes à la fin d'un chapitre ?
  - Quand le référent n'est plus mentionné pendant longtemps, la chaîne continue-t-elle pour autant ?
- Doit-on considérer des chaînes et des sous-chaînes ?
  - Quelles sont les motivations pour introduire deux niveaux de hiérarchie dans une structure a priori linéaire ?
  - Quelles sont les conséquences pratiques d'un tel choix ?
- Quelles sont les typologies des (sous-)chaînes ?
- Déjà paru :
  - Capin D., J. Glikman, C. Schnedecker & A. Todirascu (2021) « Le rôle des chaînes de référence dans la structuration textuelle : étude diachronique de l'ancien français au français moderne », *Langages n° 224*, p. 87-107

Méthodologie

- Sous-corpus de Democrat
  - 7 textes à dominante narrative du 12e au 20e
  - 13e et 14e siècles : surtout des vies de Saints

	siècle	tokens	mentions	densité réf.	Référents	singletons	CR de 3 maillons ou plus	CR de 10 maillons ou plus
Enéas	12	11643	4138	0.35	1560	1436	104	43
Jehan	15	11956	3980	0.33	1584	1466	87	36
Pantagruel	16	13833	3467	0.25	1725	1338	170	25
Clèves	17	11600	3126	0.26	1154	877	133	32
Cyrus	18	11572	3516	0.30	1507	1156	151	33
Ventre	19	12770	3147	0.24	1523	1194	140	21
Némoville	20	12660	3252	0.25	1010	724	131	32

Caractéristiques globales

	Longueur des chaînes					Distance intermaillonnaire	
	Longueur (brutes)	3 maillons et plus		10 maillons et plus			
		Moyenne	Médiane	Moyenne	Médiane	Moyenne	Médiane
Enéas	de 3 à 562	25.59	6	55.02	28	111,52	7
Jehan	de 3 à 406	28.27	12	61.11	30	97,09	8
Pantagruel	de 3 à 418	9,98	3	43.92	15	173,96	9
Clèves	de 3 à 563	14.84	4	47.50	23	182,62	12
Cyrus	de 3 à 263	13.12	4	44.78	18	157,81	10
Ventre	de 3 à 425	11.25	4	50.61	15	236,23	11
Némoville	de 3 à 299	16.93	4	56.18	27	149,30	12

→ Caractéristique des textes narratifs: beaucoup de singletons, peu de chaines longues, correspondant globalement aux personnages et lieux principaux

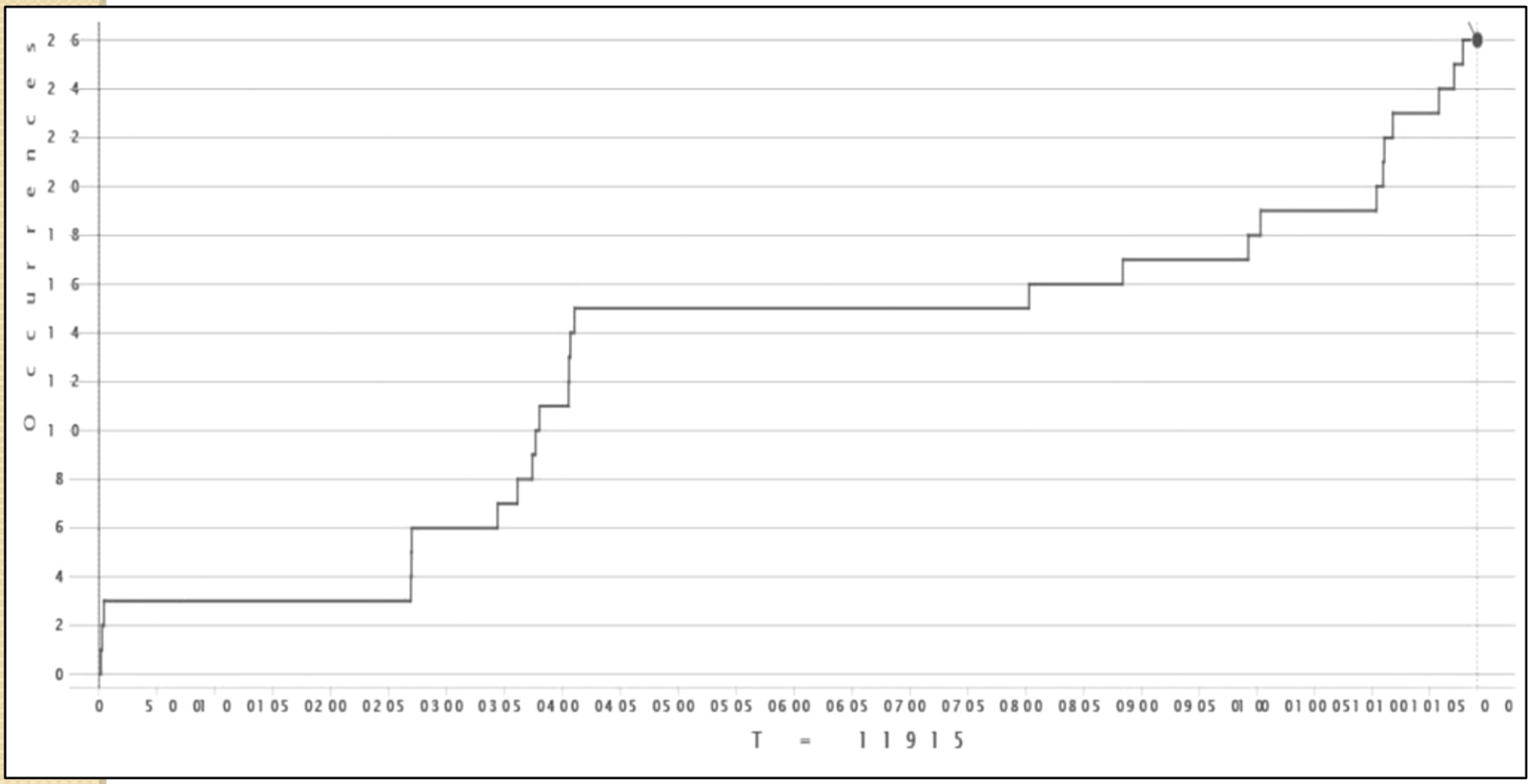
Limites des mesures quantitatives

- Longueur des chaînes de référence (CR) : Nombre de maillons, mais aussi nombre de mots couverts dans le texte

- Empan :
  Longueur CR + distance intermaillonnaire
  - Mais répartition au sein du texte ?
  - Contextes locaux différents (ex : rafales) ?
  - Lien avec la structuration ?

Visualisation de la progression

- Progression CR de Dieu dans Jehan (TXM) : une 1ère réponse
  - Peu d'occurrences
  - Mais réparties sur l'ensemble du texte



Fonction des Noms Propres (et des SN pleins)

Redénomination par des NPr ou SN

- Introduction de personnages
  - Les aultres croissoyent en long du corps : & de ceulx la sont venuz les geans, & par eulx Pantagruel Et le premier fut Chalbroth, Qui engendra Sarabroth, Qui engendra Faribroth, Qui engendra Hurtaly, qui fut beau Qui engendra Athlas, qui avecques ses espaulles garda le ciel de tumber, Qui engendra Goliath, Qui engendra Eryx lequel fut inventeur du jeu des gobeletz, [...] Qui engendra Grand Gosier Qui engendra Gargantua, Qui engendra le noble Pantagruel mon maistre (Pantagruel)
- Désambiguisation référentielle
  - Ung jour comme **le roy**$_{RFr}$ venoit de la messe acompaigné de **ses**$_{RFr}$ barons et chevaliers, et ainsi qu'**il**$_{RFrance}$ estoit à l'entree de **son**$_{RFr}$ palaix royal, et que celuy jour estoit une solempne feste, arriva devant **luy**$_{RFr}$ **le roy d'Espaigne**$_{REsp}$ **qui**$_{REsp}$ en grans pleurs et gemissemens se gecta aux piez **du roy de France**$_{RFr}$. (Jehan)

Fonction des Noms Propres

Rôle dans la structuration textuelle

- Avec d'autres marques, dont paragraphe :
  - (§ 6) **Le roi de Navarre** attirait le respect de tout le monde par la grandeur de son rang et par celle qui paraissait en sa personne. Il excellait dans la guerre ; et le duc de Guise lui donnait une émulation qui l'avait porté plusieurs fois à quitter sa place de général pour aller combattre auprès de lui, comme un simple soldat, dans les lieux les plus périlleux. […]

    (§ 7) **Le roi** avait ***toujours*** aimé le connétable ; et sitôt qu'il avait commencé à régner, il l'avait rappelé de l'exil où le roi François Ier l'avait envoyé. […] (Clèves)

- mais aussi marqueurs spatiaux, temporels…

Fonction des Noms Propres

- Sans autre marque : redénomination marque à elle seule un changement de séquence textuelle (descriptif > narratif), de focalisation…

  (§ 4) (...) **Elle** (= Madame François) s'était penchée, **elle *avait aperçu***, à droite, presque sous les pieds du cheval, une masse noire qui barrait la route.

  (§ 5) — On n'écrase pas le monde, dit-**elle**, en sautant à terre.

  (§ 6) <<<SEQ. DESCRIPTIVE>>> **C'**était un homme vautré tout de son long, les bras étendus, tombé la face dans la poussière. Il paraissait d'une longueur extraordinaire, maigre comme une branche sèche ; le miracle était que Balthazar ne l'eût pas cassé en deux d'un coup de sabot. <<< CONJECTURE$_1$>>>> **Madame François** le ***crut*** mort ; **elle s**'accroupit devant lui, **ø** lui prit une main, et ***ø vit*** qu'elle était chaude.

  (§ 7) <<<VERIFICATION CONJECTURE$_1$>>>> — Eh ! l'homme ! dit-**elle** doucement.

  (§ 10) Cependant, *l'homme* avait ouvert les yeux. *Il* regardait **madame François** d'un air effaré, sans bouger. **Elle** pensa qu'*il* devait être ivre, en effet. (*Ventre*)

- Fonction de structuration dans les textes anciens sans « paragraphe » : narration plate

CONCLUSION ET PERSPECTIVES

Apport du corpus Democrat

- Possibilités d'études à la fois quantitatives et qualitatives
- Dans des textes comparables
- Avec des outils intégrés (macros TXM)
- Permet des études linguistiques aussi bien que des applications de traitement automatique des langues
- Représente la langue française dans des ressources internationales orientées sur anaphores et coréférences
- Intégré à *Universal Anaphora* (*Coref-UD, Coreference Universal Dependancies*) initiative visant à constituer un grand corpus multilingue pour les annotations d'anaphores
- Présent dans la campagne internationale "CRAC 2022" Shared Task on Multilingual Coreference Resolution

Bilan du projet Democrat

- Mise à disposition à une large communauté de données enrichies et de nouvelles connaissances sur la langue : FAIT, uniquement pour les expressions référentielles et les chaînes de référence
- Mise à disposition de nouveaux outils et de nouveaux procédés de visualisation pour la manipulation de ces données et connaissances : FAIT, initialement pour les expressions référentielles et les chaînes, mais potentiellement pour des objets beaucoup plus généraux
- Réflexion sur l'annotation des chaînes de référence et des structures textuelles : abordée, mais pas mise en œuvre
- Réflexion sur l'inter-opérabilité et donc sur une exploitation croisée des annotations des chaînes de référence et des structures textuelles : abordée, mais pas mise en œuvre

Perspectives de Democrat

- Revenir du corpus et de l'analyse de ses annotations à l'élaboration d'un modèle linguistique discursif

- Democrat comporte trois variations :
  - Genre textuel – variation matérialisée dans le corpus
  - Date : approche diachronique – matérialisée dans le corpus
  - Langue : approche contrastive – non matérialisée dans le corpus

- D'autres variations sont envisageables
  - Productions de sujets pathologiques (psycholinguistique)
  - Ecrit versus oral, voire nouvelles formes de communication

- A plus long terme, le travail de Democrat pourrait être un premier pas vers des recherches sur les aspects cognitifs de la référence, avec notamment la notion de saillance