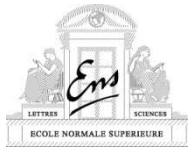
Étude des chaînes de référence en français :

le projet ANR Democrat

Frédéric Landragin

Séminaire CLLE-ERSS

19 novembre 2020



Contenu

- Problématiques et objectifs
- Un cadre de travail : le projet ANR Democrat

1. La référence, les expressions référentielles, les chaînes de coréférences
2. Le corpus Democrat : constitution, annotation, double annotation
3. Linguistique outillée pour l'analyse des références et des chaînes de coréférences
4. Traitement automatique des langues : détection automatique des chaînes de coréférences

- Bilan et perspectives

Problématiques et objectifs

Trois objets d'étude

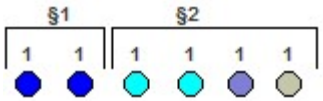
Comme tout avait brûlé – la mère, les meubles et les photographies de la mère -, pour Fabre et le fils Paul c'était tout de suite beaucoup d'ouvrage : toute cette cendre et ce deuil, déménager, courir se refaire dans les grandes surfaces. Fabre trouva trop vite quelque chose de moins vaste, deux pièces 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.

Suite des références :

1 = Sylvie Fabre 2 = Mr. Fabre 3 = Paul Fabre
4 = {Mr. Fabre, Paul Fabre}

Chaîne de coréférences qui concerne Sylvie Fabre :



la mère – la mère – sa mère – sa mère – Sylvie Fabre – la

Définitions et objectifs

- Références : accès *vs.* évocation d'un référent
  - **expression** linguistique qui réfère à un référent, et fait de ce référent une entité du discours impliquée dans les structures syntaxique, sémantique, informationnelle de la phrase
  - **indice** linguistique (morphème, sujet zéro) qui évoque un référent, sans référer, tout en contribuant à la saillance du référent
- Suite des références
  - succession des expressions référentielles
  - étude des transitions d'un référent à un autre, anaphores associatives…
  - vers une typologie des transitions référentielles : continuation sur un même référent, bifurcation, confrontation de deux référents…
- Chaînes de coréférences
  - succession des expressions et indices qui concernent un même référent
  - étude des typologies de chaînes

Questions sous-jacentes

- Nature des expressions référentielles
  - qu'est-ce qui réfère dans un texte ?
  - qu'est-ce qui évoque un référent sans pour autant référer ?
  - si on distingue plusieurs degrés de référence, comment en tenir compte dans une méthodologie d'annotation de corpus ?
- Nature des chaînes de coréférences et liens avec la suite des références
  - comment une chaîne commence-t-elle ? se termine-t-elle ?
  - quelles sont les typologies des chaînes ?
  - comment les chaînes se croisent-elles dans un texte ?
  - peut-on prévoir des motifs dans la succession des références ?
  - quelles sont les corrélations entre typologies des chaînes et données syntaxiques, sémantiques, pragmatiques ?
  - comment définir de manière opérationnelle la saillance ?

Premières étapes de travail

- Identifier et catégoriser les référents (objets du monde)
- Relier entre eux les référents (groupes et individus)
- Identifier et catégoriser les expressions référentielles
- Relier entre elles les expressions référentielles, c'est-à-dire construire les chaînes de coréférences
- Caractériser les chaînes de coréférences

Enjeux scientifiques

- Proposer un modèle « intégré » de la référence
  - qui étudie la référence et la coréférence du point de vue du discours et pas uniquement localement
  - qui s'enrichit de comparaisons avec d'autres langues (approche contrastive) et avec plusieurs états de langue (approche diachronique)
  - qui tient compte du genre textuel
- Faire un pont entre théories linguistiques et techniques de traitement automatique des langues
  - annoter un corpus en vue de fournir des données d'apprentissage
  - souligner les phénomènes négligés jusque-là par le traitement automatique des langues
- Proposer le premier système de détection automatique *end-to-end* pour la langue française

Apports et retombées escomptés

- Mise à disposition à une large communauté de données enrichies et de nouvelles connaissances sur la langue
- Mise à disposition de nouveaux outils et de nouveaux procédés de visualisation pour la manipulation de ces données et connaissances
- Mise à disposition de nouvelles méthodes d'analyse (linguistique et statistique) des chaînes de coréférences
- Représentation de systèmes de TAL traitant la langue française dans des campagnes d'évaluation internationales
- Contribution aux humanités numériques
  - à la pérennisation des données, à leurs standardisation,
  - à la place du français dans le monde,
  - à la didactique et à l'enseignement du français et des langues

Un cadre de travail : Le projet ANR Democrat

https://www.lattice.cnrs.fr/democrat/

A l'origine : un groupe de travail Lattice appelé COREF

novembre 2008	définition des objectifs du groupe de travail
décembre 2008	expressions référentielles et ambiguïtés
janvier 2009	référents évolutifs ; groupes stricts et flous
mars 2009	méthodologie d'annotation des références
mars 2009	relations entre référents et Théorie des Ensembles Flous
avril 2009	typologie des transitions référentielles ; métaphore cinéma
juin 2009	types d'introduction de référents ; MMAX vs. GLOZZ
novembre 2009	motifs pour repérer les transitions référentielles
décembre 2009	méthodologie d'annotation, schéma d'annotation
janvier 2010	interactions entre individus et autres entités
janvier 2010	*séance spéciale TEI, ANANAS…*
janvier 2010	***atelier Lattice : premier bilan***

Planning du groupe COREF

février 2010	Théorie du Centrage ; annotation de la saillance
mars 2010	chaînes de coréférence comme marqueurs de thème ; GLOZZ
avril 2010	pluralité, groupe, collectif, collection ; évocation de référent
mai 2010	étude contrastive français-hongrois ; la définitude en diachronie
juin 2010	un seul centre du discours vs. plusieurs échelles de saillance
septembre 2010	coréférence et TAL ; pronoms, prédications, attributions ; ANALEC
octobre 2010	projet COREF ; Labex ; ambiguïtés et sous-déterminations
décembre 2010	portée d'une chaîne ; références aux entités non humaines
janvier 2011	noms de fonctions et expressions attributives
janvier 2011	*séance spéciale TAL : méthodes, algorithmes, évaluation, projets*
février 2011	schéma d'annotation « niveau 0 » ; étiquettes et coréférences
mars 2011	défini vs. démonstratif ; maillons forts et faibles d'une chaîne
mars 2011	***atelier Lattice : deuxième bilan***

Puis un projet PEPS : MC4

Puis un projet ANR : Democrat

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

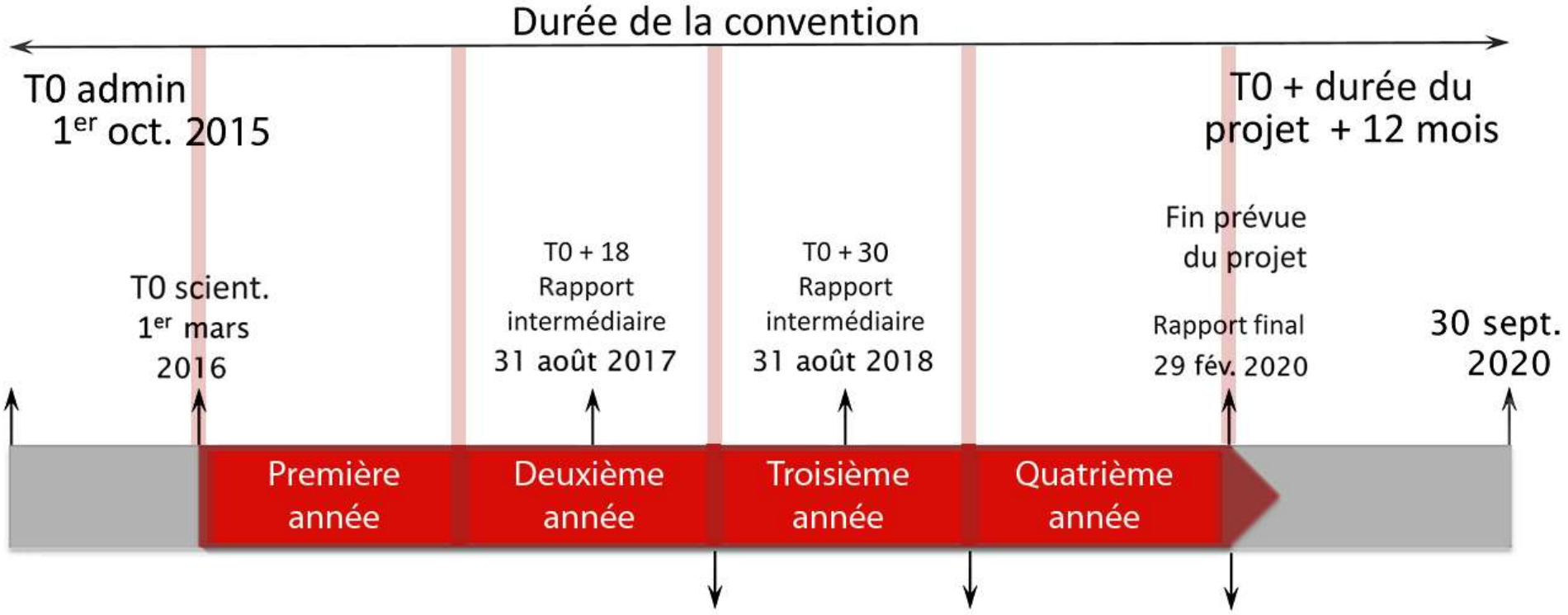
Participants du projet

- Partenaire 1 : ENS Paris, laboratoire Lattice
  - Responsable : Frédéric Landragin, coordinateur du projet
  - 15 participants : 10 membres du Lattice, 5 personnes rattachées d'autres laboratoires, des doctorants (dont un CDD doctoral financé par Democrat) et un post-doc (CDD d'un an financé par Democrat)
- Partenaire 2 : Université Strasbourg, laboratoire LiLPa
  - Responsable : Catherine Schnedecker
  - 13 participants : plusieurs membres du LiLPa, 3 personnes rattachées d'autres laboratoires, pas mal de doctorants et un post-doc (CDD d'un an financé par Democrat)
- Partenaire 3 : ENS Lyon, laboratoires ICAR et IHRIM
  - Responsable : Céline Guillot-Barbance
  - 9 participants : 3 membres d'ICAR, 4 d'IHRIM dont un ingénieur (CDD de deux ans financé par Democrat)

4 ans – 4 objectifs – 4 livrables

1. Modélisation linguistique (discursive, contrastive…)
   pas de livrable spécifique si ce n'est le pseudo-livrable « publications et formations », commun aux 4 objectifs et étalé sur les 4 ans du projet
2. Constitution d'un corpus annoté
   livrable « méthodologie d'annotation » : livré en mars 2018, qui amènera au livrable « corpus » livré en mai 2019
3. Conception d'un outil d'annotation et d'interrogation
   livrable « TXM » livré en mai 2019
4. Conception d'un système de détection automatique
   livrable « TAL », avec deux systèmes réalisés selon différentes approches et préoccupations, livré en mars 2020

Planning du projet



Workshops du projet

- Juin 2015 : journée d'étude « les chaînes de référence en corpus » au LiLPa
- Mars 2016 : réunion plénière « kick-off » au Lattice (2 jours)
- Février 2017 : réunion plénière au Lattice (2 jours)
- Novembre 2017 : journée d'étude « chaînes de référence et structures textuelles » à l'ENS de Lyon
- Mars 2018 : réunion plénière au Lattice (2 jours)
- Mars 2018 : journée d'étude « approches contrastive des chaînes de référence » au Lattice
- Mars 2019 : réunion plénière au Lattice
- Juin 2019 : journée d'étude « mesures pour étudier les chaînes de référence » au LiLPa
- Février 2020 : réunion de clôture au Lattice

Publications collectives marquantes

La référence,
les expressions référentielles,
les chaînes de coréférences

Volet 1 « linguistique » du projet Democrat
https://www.lattice.cnrs.fr/democrat/publications.html
https://www.hal.archives-ouvertes.fr/DEMOCRAT/

Problèmes de définition

- « Le village était désert. Il semblait abandonné.
  La place principale était vide. Elle en paraissait triste.
  Tout reprendrait vie le lendemain matin, Ø repartirait de zéro :
  le village s'animerait, la place se remplirait de monde »
- Quels sont les référents ? « en » a-t-il un référent ?
- Quelles sont les expressions référentielles ?
  sur quels critères intégrer les sujets non exprimés ?
- Quelles sont les chaînes de références ? de coréférences ?
  les chaînes anaphoriques ?
- Quels sont les antécédents ? première ou dernière expression mentionnée ?

Problèmes liés à la détermination des expressions référentielles

- La référence est un problème linguistique, qui a des conséquences sur la procédure d'annotation
- Analyse d'un autre exemple :
  - « Pierre et Paul ont chacun eu un fils cette année.
    Il se trouve qu'ils ont la même nourrice. »
  - personnages : Pierre, le fils de Pierre, Paul, le fils de Paul, la nourrice
  - « Pierre et Paul » : du fait de la coordination, faut-il considérer qu'il y a référence à un groupe d'individus ?
  - « un fils » : est-ce référentiel ?
  - « ils » : réfère apparemment au groupe des deux fils, sauf que ce groupe n'a pas été évoqué précédemment. Est-ce pour autant une première mention (ie. le premier maillon de la chaîne) ?
  - « chacun » ?

Formes pleines et formes atténuées

- En plus des expressions référentielles, certains mots ou morphèmes participent aux chaînes de coréférences
  - les marques d'accord en genre et/ou en nombre (qui, sans référer, rappellent le référent et participent ainsi aux coréférences) :
    - dans « ils dorment », la terminaison « -ent » rappelle le pluriel
    - à annoter ? via une catégorie spécifique ?
  - les sujets zéro (infinitifs, participiales…) :
    - l'intérêt de les annoter est qu'ils peuvent être saillants et contribuer ainsi fortement aux coréférences
    - on peut alors confronter des exemples tels que « il entra, il prit son chapeau » et « il entra, prit son chapeau »
    - à annoter ? comme on n'annote pas du vide (ni un signe de ponctuation), il faut recourir à une solution telle qu'annoter le verbe en tant que support d'une expression référentielle élidée
  - les constructions pronominales, etc.

Le cas des attributs et étiquettes

Je suis sursitaire, âgé de 24 ans, et je suis marié à une veuve de 44 ans, laquelle a une fille qui en a 25. Mon père a épousé cette fille. A cette heure, mon père est donc devenu mon gendre, puisqu'il a épousé ma fille[1]. De ce fait, ma belle-fille[2] est devenue ma belle-mère, puisqu'elle est la femme de mon père.

Ma femme et moi avons eu en Janvier dernier un fils. Cet enfant est donc devenu le frère de la femme de mon père, donc le beau-frère de mon père. En conséquence, mon oncle, puisqu'il est le frère de ma belle-mère. Mon fils est donc mon oncle.

[1] Il manque l'étape : « la fille de ma femme » devient « ma fille »…

[2] On ignore le référent indirect, de même que dans « un parricide »

- Certaines expressions réfèrent, d'autres servent d'étiquettes

Problèmes de délimitation d'une expression référentielle

Quelques exemples de premières mentions :

1. **Le président Jacques Chirac** a dit…
2. **Le président de la République, Jacques Chirac**, a dit…
3. **Jacques Chirac, président de la République,** a dit…
4. **Jacques Chirac** – eh oui ! – **président de la République,** a dit…
5. **Le président de la République, qui s'appelle Jacques Chirac,** a dit…
6. **Cet imbécile de président** a dit…
7. **Jacques Chirac** est **le premier président qui a été maire de Paris**.

Plusieurs possibilités selon les cas :

- une seule expression référentielle (qui groupe parfois plusieurs syntagmes)
- plusieurs expressions référentielles
- plusieurs expressions, seule la 1ère étant considérée comme référentielle
- plusieurs expressions, la plus directe (nom propre) étant considérée référentielle

Problèmes d'annotation :

- on a parfois du mal à déterminer des limites précises → borne inf et borne sup
- l'exemple avec du texte discontinu pose des problèmes techniques

Attribuer un référent peut s'avérer impossible

- Certains pronoms peuvent rester ambigus, même en tenant compte des connaissances encyclopédiques d'un lecteur averti
- Exemple : résumé du film « *Le cave se rebiffe* »

  Eric Masson, un "demi-sel", est devenu l'amant de la belle Solange Mideau, femme d'un graveur raté. Eric veut se servir de Robert Mideau pour monter, à **son** insu, un trafic de fausse monnaie. Il s'associe à Charles Lepicard, tenancier d'une ancienne maison close, et à Lucas Malvoisin, l'homme d'affaires de celui-ci. Charles et Lucas n'ont pas grande confiance en Eric, mais Solange **leur** promet **son** concours. Elle souhaite en effet mener la grande vie. Avec l'accord de **ses complices**, Charles contacte Ferdinand Maréchal, dit le Dabe, vieux truand célèbre qui s'est retiré dans une île des Tropiques. Il le décide à venir à Paris.

  - « à son insu » : Robert Mideau ou Solange Mideau ? pas si simple…
  - « leur » : Charles (sûr) + Lucas (sûr) + Eric (peu probable, mais possible)
  - « son concours » : ambigu entre Solange et Eric
  - « ses complices » : Lucas (sûr) + Solange (pas si sûr) + Eric (?)
  - question subsidiaire : qui est « le cave » ?

    → rôle du titre ?

Attribuer un référent peut évoluer au fur et à mesure de la lecture

- Il arrive qu'en poursuivant la lecture d'un texte, on remette en question une référence a priori non ambiguë

L'ancien président de la République de Côte d'Ivoire, Henri Konan Bédié et son épouse ont reçu à dîner l'ancien Premier ministre Alassane Dramane Ouattara et son épouse, le 23 septembre. La rencontre très médiatisée avait un objectif, celui de montrer que **les héritiers du premier président de Côte d'Ivoire** peuvent se retrouver pour reconquérir le pouvoir. **Les deux leaders** ont l'habitude de se voir et de s'appeler depuis le déclenchement, le 19 septembre 2002 de la rébellion en Côte d'Ivoire. A Paris, à Abidjan, à Accra, **les deux hommes** se côtoient, mais dans des cadres formels. **Leur** rencontre en soi n'est donc pas un événement, sauf qu'**ils** ont voulu donner à cette entrevue un cachet particulier. Les retrouvailles autour d'un même idéal politique que commande la mémoire du "Vieux" dont **ils** se réclament. […]
Mais après que tout le monde ait perdu le pouvoir, en faveur d'**un autre héritier, le général Robert Guéi**, par un coup d'Etat en décembre 1999, la gestion du pays semble échapper aux "**enfants**".

  - au début : « les héritiers » = H.K.B. + A.D.O.
  - il y a du flou : « les héritiers » = H.K.B. + A.D.O. + leurs femmes
  - puis : « un autre héritier » = R.G., d'où :
    - nécessairement, « les héritiers » = H.K.B. + A.D.O. + R.G. + ?
    - et, finalement : « les héritiers » = groupe de personnes aux limites floues, qui comprend au moins les trois hommes cités

Conséquences sur l'annotation : plusieurs stratégies sont possibles

1. **On se focalise sur les formes linguistiques**, sans tenir compte des éventuelles ré-interprétations ultérieures (stratégie linéaire)
   - avantage : théoriquement, on réduit les biais interprétatifs et on rend compte des étapes de l'interprétation
   - inconvénients : vouloir attribuer un référent sur la seule forme linguistique est illusoire, car nos connaissances encyclopédiques interviennent constamment ; annoter quelque chose de faux n'est pas très pertinent…
2. **On se focalise sur les concepts**, et on n'annote qu'après avoir compris tout le texte et calculé toutes les références
   - avantage : on se rapproche de la réalité modélisée derrière le texte
   - inconvénient : on s'éloigne des effets stylistiques voulus par l'auteur
3. **On part des concepts et on élargit aux interprétations possibles**, via un attribut dédié (interprétation immédiate vs. différée)
   - avantage : on modélise de manière complète
   - inconvénient : rédiger un manuel d'annotation peut s'avérer compliqué

Conséquences sur l'annotation : c'est une opération floue…

- On ne tente pas donc d'attribuer un référent à tout prix, mais on prend en compte les possibilités d'ambiguïté, d'imprécision, de flou
- On prend en compte la notion de flou, d'une part pour la détermination des groupes (groupe strict versus groupe flou), d'autre part pour la relation d'appartenance à un groupe (appartenance stricte versus floue)
- On peut modéliser ces aspects avec la Théorie des Ensembles Flous (Zadeh)
  - « Solange Mideau » (référence individuelle sans problème) : A_{strict}
  - « son concours » (ambigu entre Solange et Eric) : A_{strict} **ou** B_{strict}
  - « le cave » : A_{flou}
  - « Charles et Lucas » (groupe construit) : $groupe_{strict}\ \{ A_{strict} ; B_{strict} \}$
  - « ses complices » : $groupe_{strict}\ \{ A_{strict} ; B_{strict} ; C_{flou} \}$
  - « les héritiers » : $groupe_{flou}\ \{ A_{strict} ; B_{strict} ; C_{flou} ; D_{flou} \}$

Le corpus Democrat : constitution, annotation, double annotation

Volet 2 « corpus » du projet Democrat
https://www.ortolang.fr/market/corpora/democrat/

Constitution du corpus

- 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 construites (dont 9 000 > 2 maillons)

Annotation du corpus

- Groupe COREF : annotations éparses, non homogènes, correspondant aux préoccupations personnelles des participants
  Bilan : inexploitable !

- Projet PEPS MC4 : annotation très complète et très fine de l'ensemble des questions linguistiques discutées
  Bilan : corpus tout petit !

- Projet ANR Democrat : annotation très simple des expressions référentielles : au minimum le référent, au mieux le référent + la catégorie d'expression référentielle
  Bilan : gros corpus, mais annotations « minimalistes »

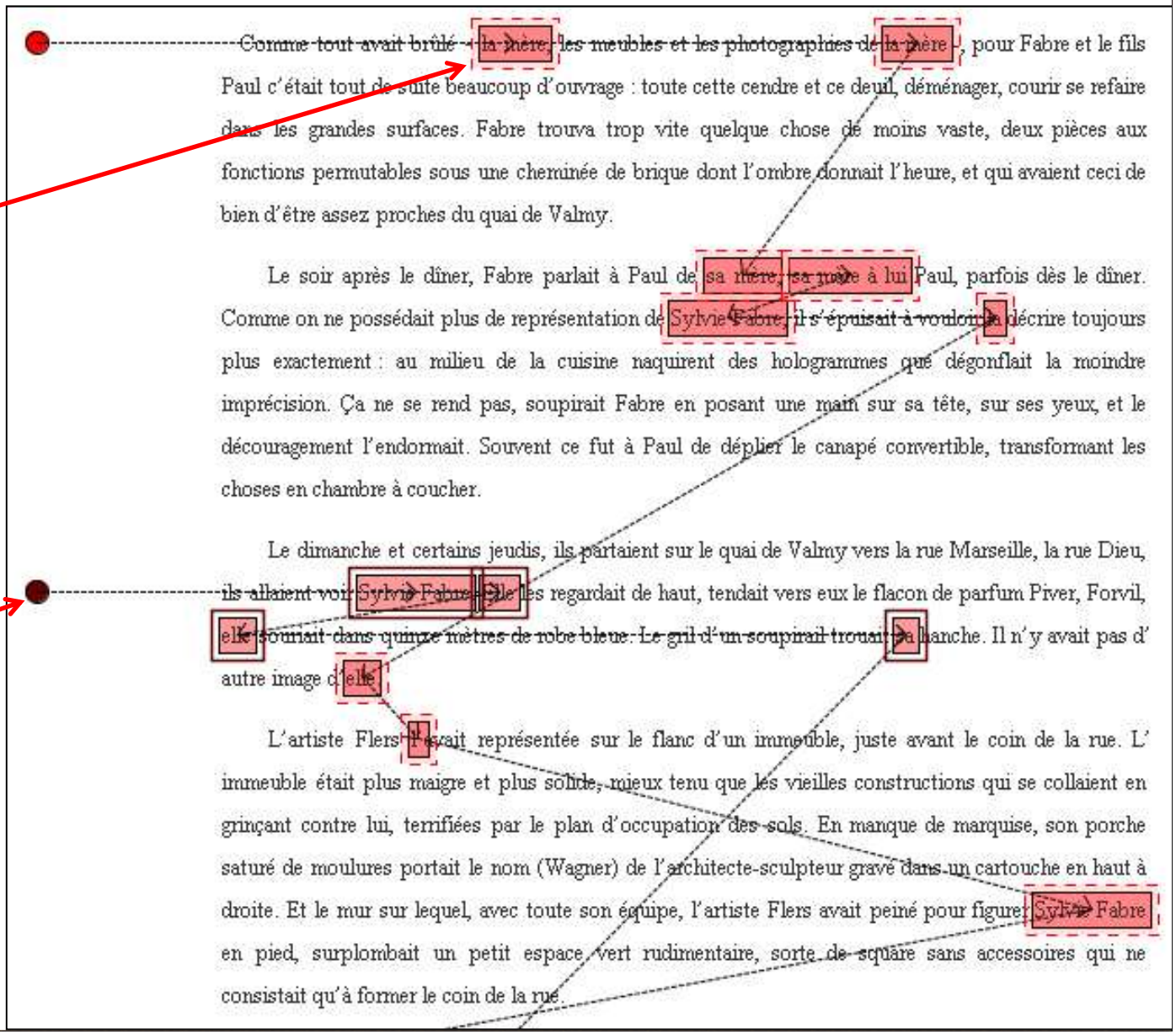
Le modèle URS de Glozz

- 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

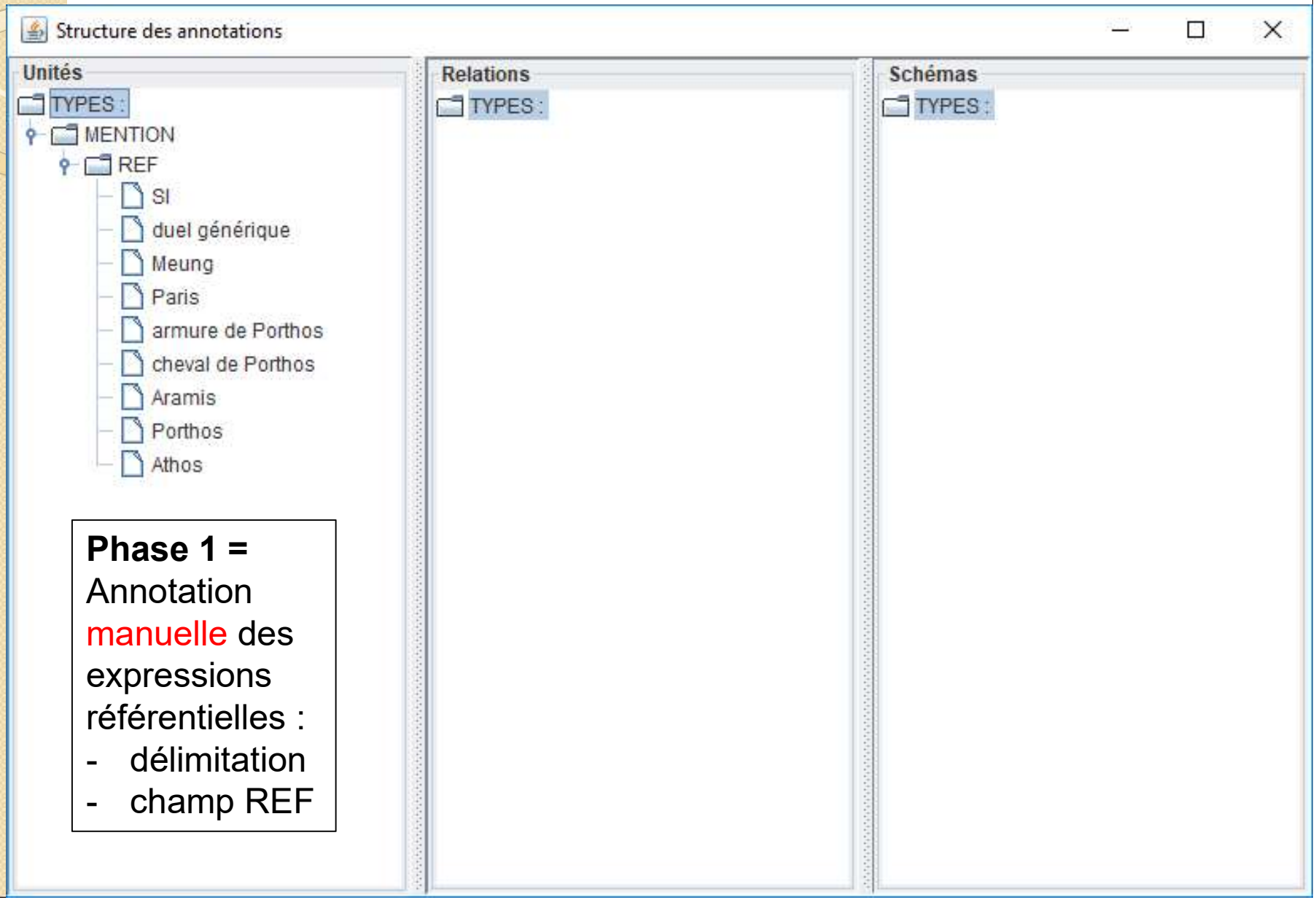
Matérialisation des choix du projet

Une unité de type « expression référentielle »

Un schéma de type « chaîne de coréférences »

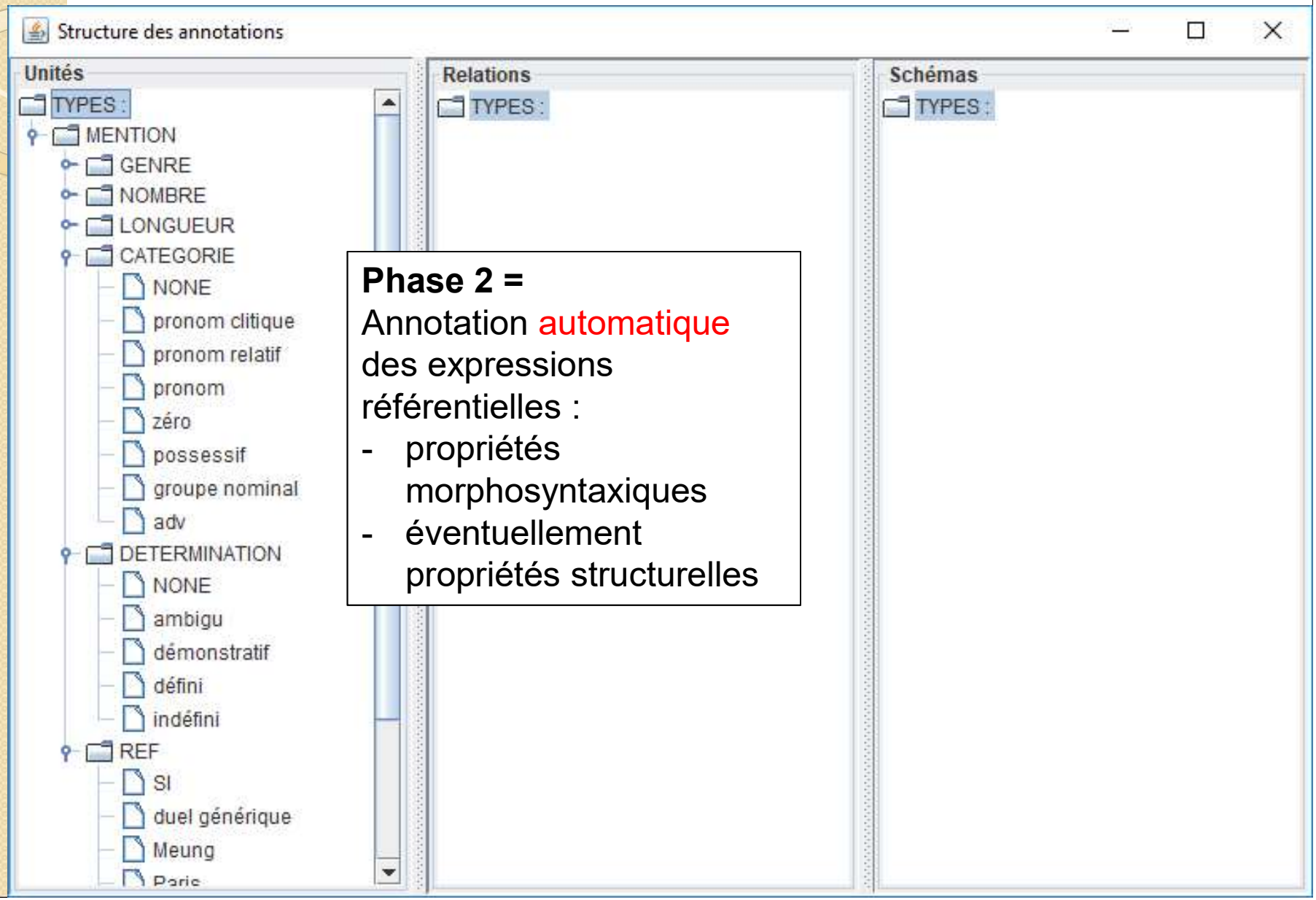


Procédure d'annotation du corpus



Phase 1 =
Annotation manuelle des expressions référentielles :
- délimitation
- champ REF

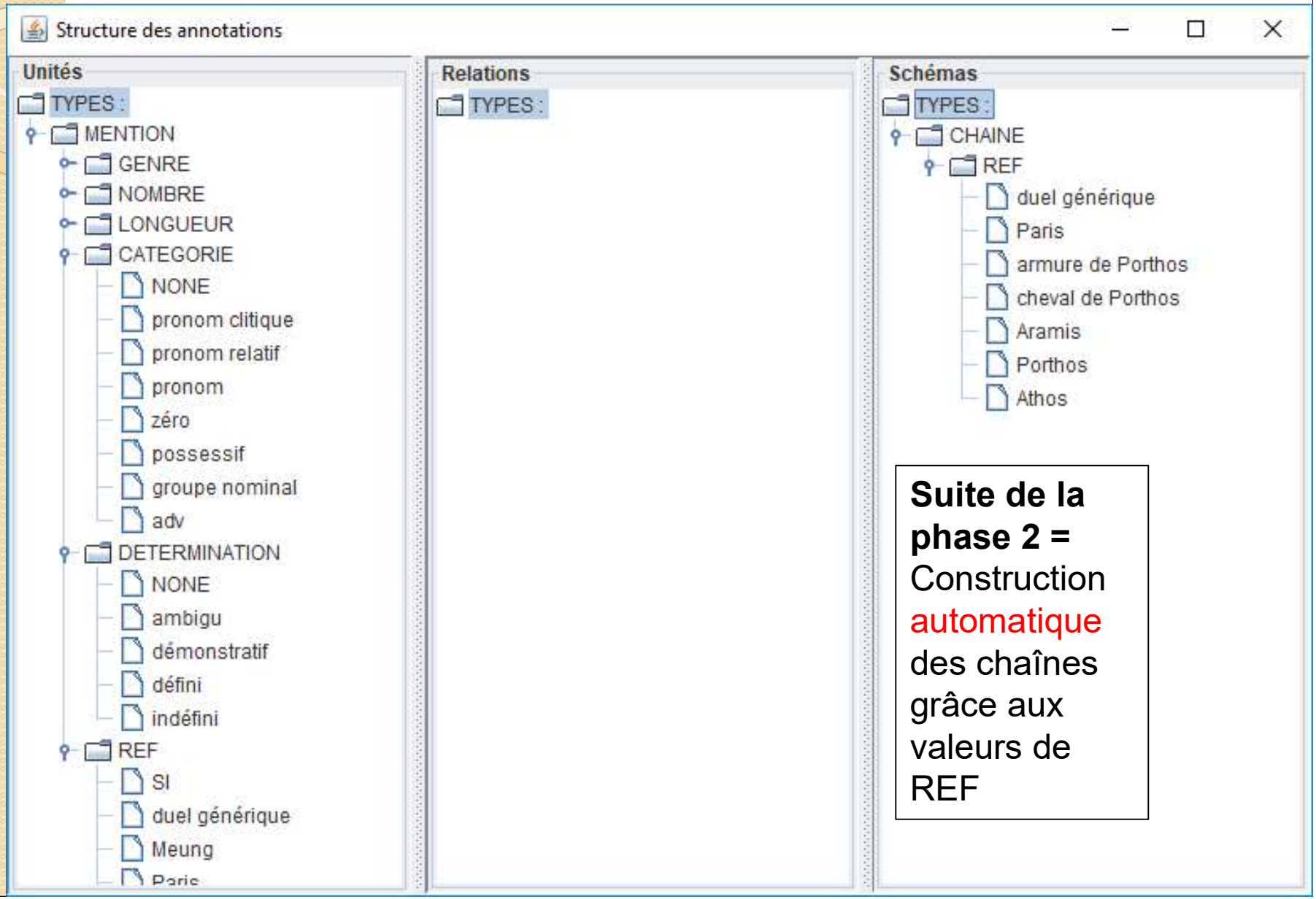
Procédure d'annotation du corpus



Phase 2 =
Annotation automatique des expressions référentielles :

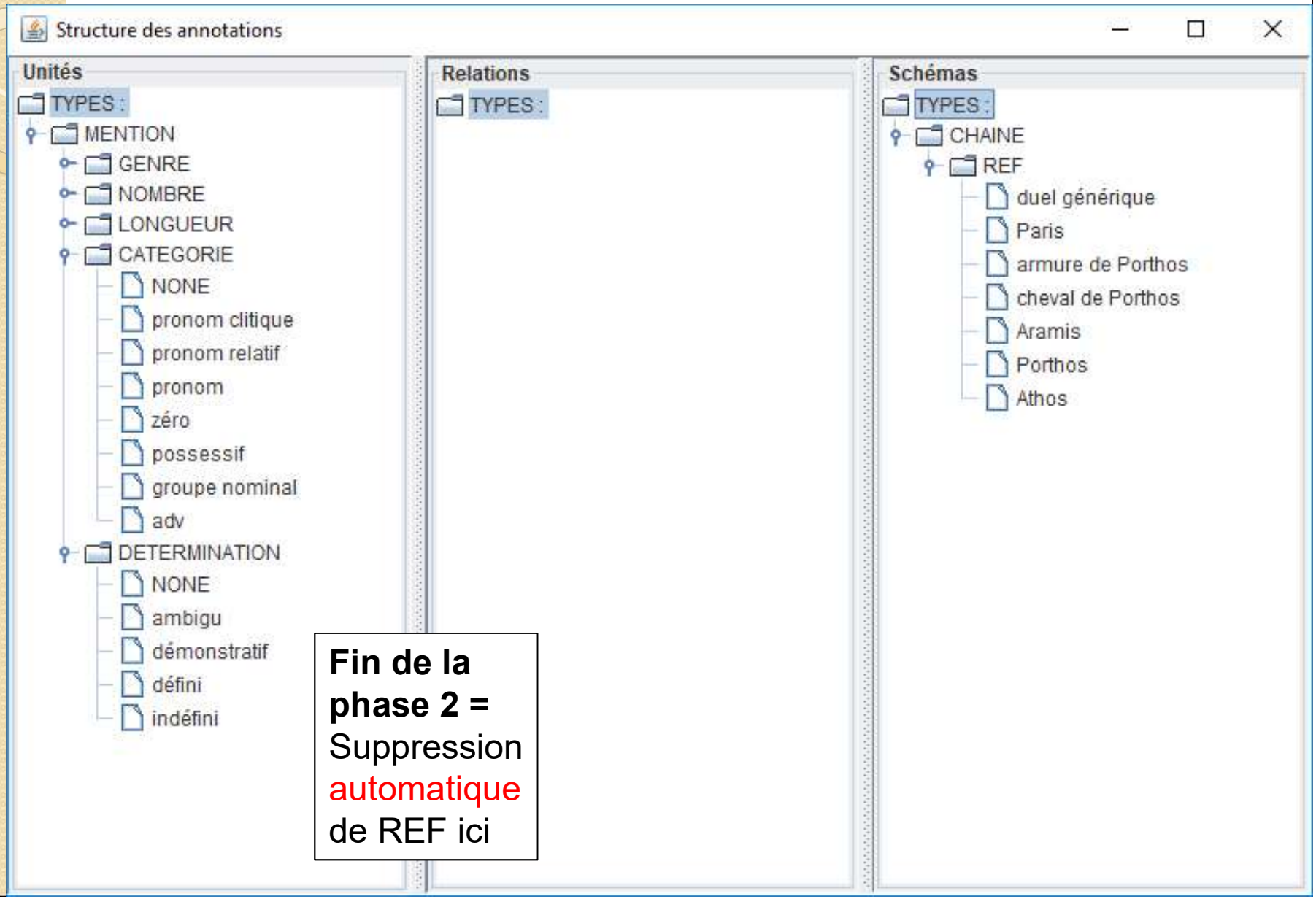
- propriétés morphosyntaxiques
- éventuellement propriétés structurelles

Procédure d'annotation du corpus



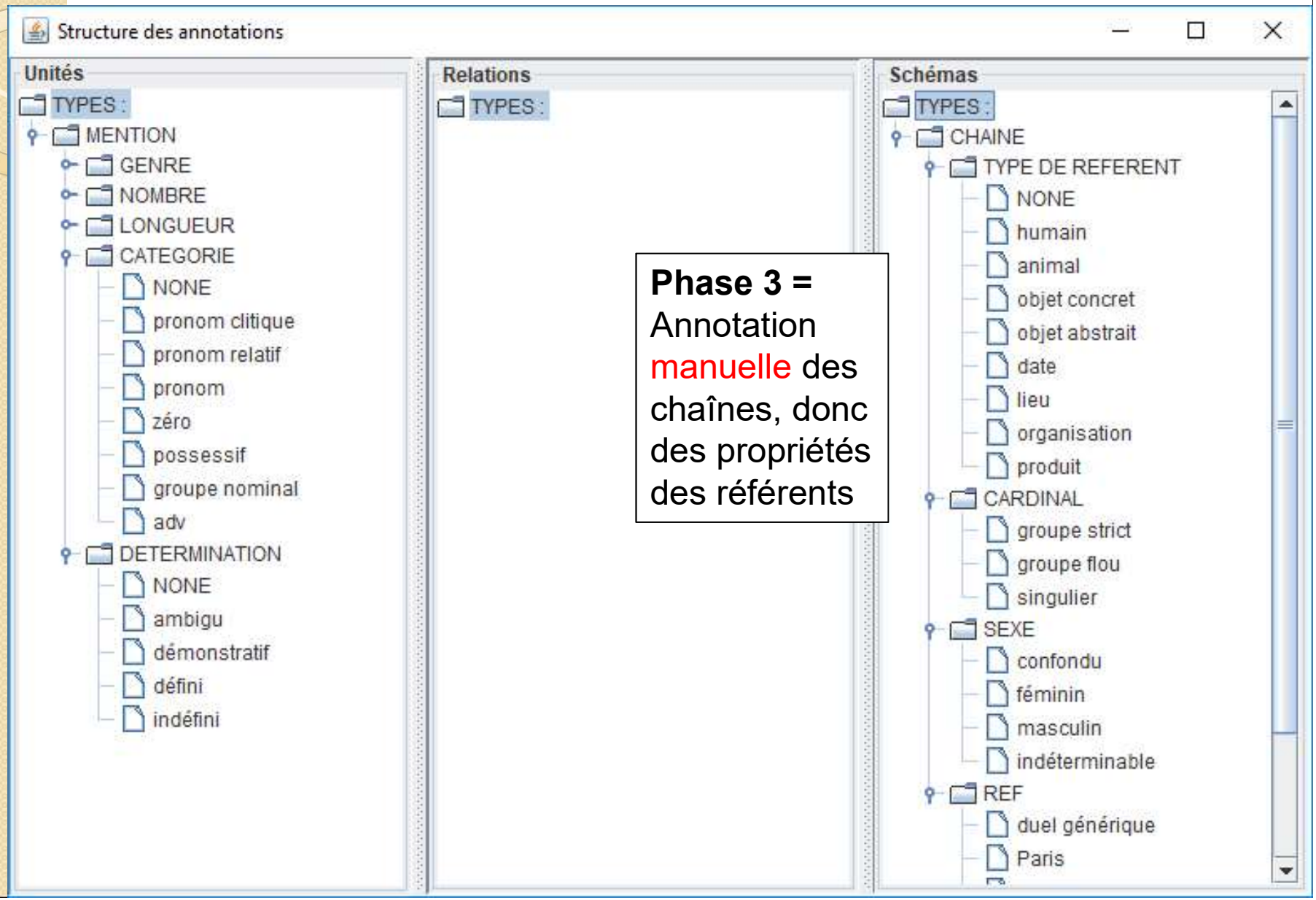
Suite de la phase 2 = Construction automatique des chaînes grâce aux valeurs de REF

Procédure d'annotation du corpus



Fin de la phase 2 = Suppression automatique de REF ici

Procédure d'annotation du corpus



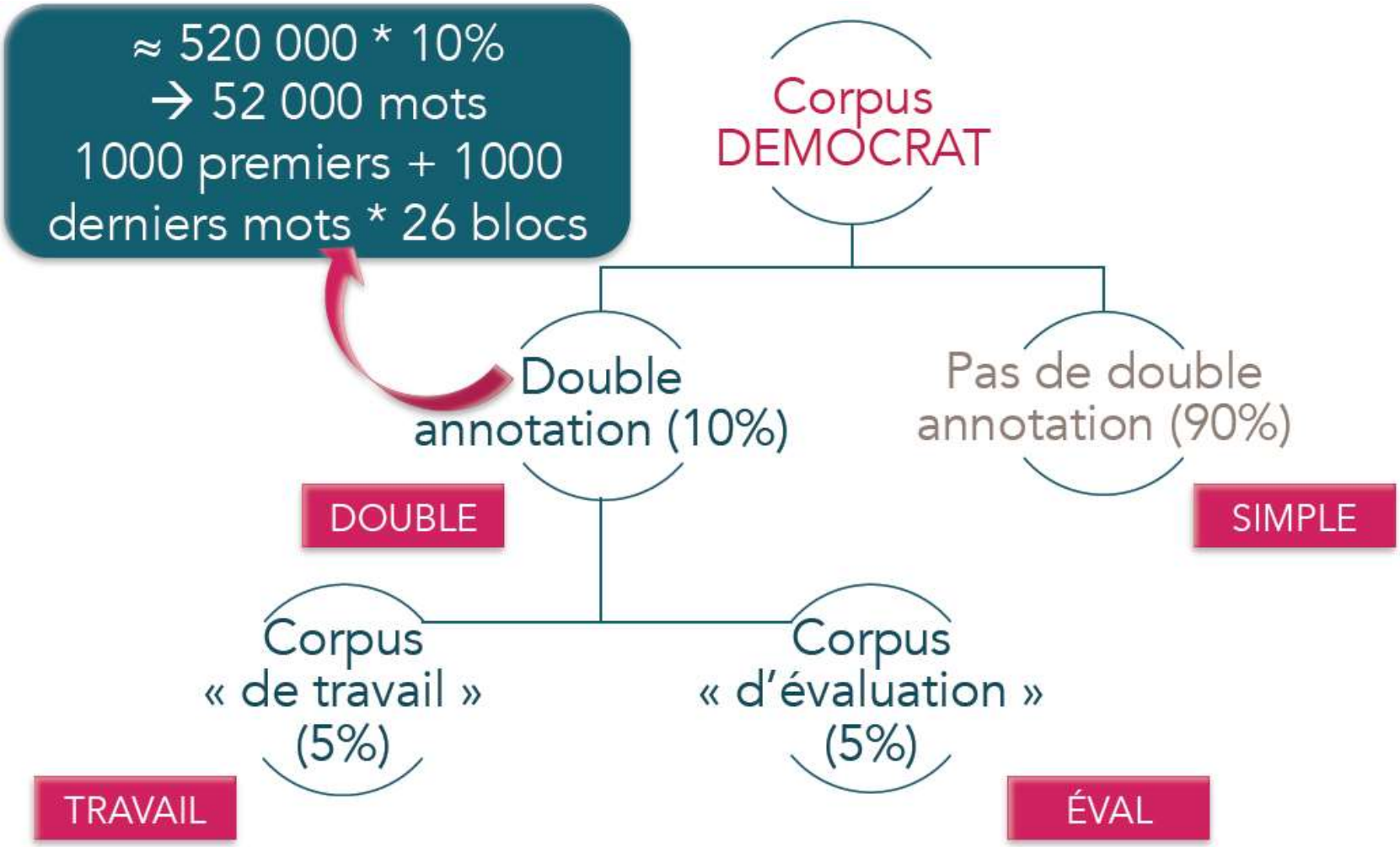
Phase 3 = Annotation manuelle des chaînes, donc des propriétés des référents

Annoter les expressions référentielles, annoter les chaînes

- La tâche la plus complexe et la plus chronophage est le repérage des expressions référentielles
  - Toutes les expressions référentielles ! Pas seulement celles qui réfèrent à des humains !
  - D'où un grand nombre de « singletons » (référents spatiaux, temporels)
  - Nombreuses difficultés pour délimiter les expressions : problèmes avec les relatives, les appositions, etc.
  - Le manuel d'annotation comporte plus de 30 pages qui décrivent tout un ensemble de cas

- La 2e tâche importante est l'affectation d'un référent à chaque expression référentielle
  - Face à une ambiguïté, on doit choisir…
  - Pas de place pour le flou ou pour une approche « good-enough »…
  - C'est ainsi que les chaînes sont construites

Evaluation de la qualité et scission du corpus



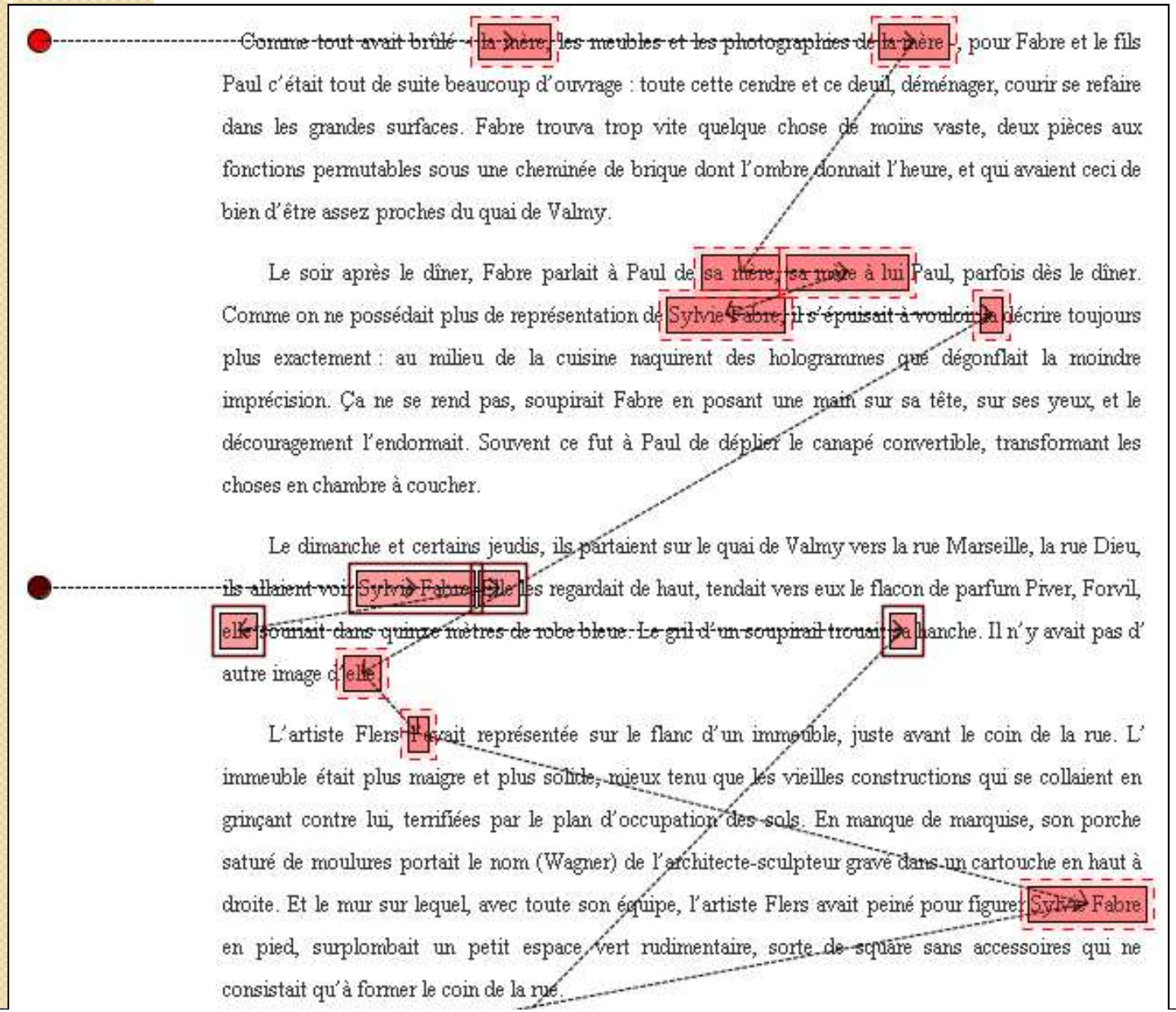
Mais, au fait, à quoi servent les annotations ?

- Constituer un corpus de référence sur la référence et la coréférence
- Fournir aux linguistes un « réservoir » d'exemples qui soit riche et varié
- Fournir des données pour des calculs statistiques voire textométriques sur les chaînes de coréférences
- Fournir des données d'apprentissage pour la mise en œuvre de système de détection automatique de chaînes de coréférences

Linguistique outillée pour l'analyse des références et des chaînes de coréférences

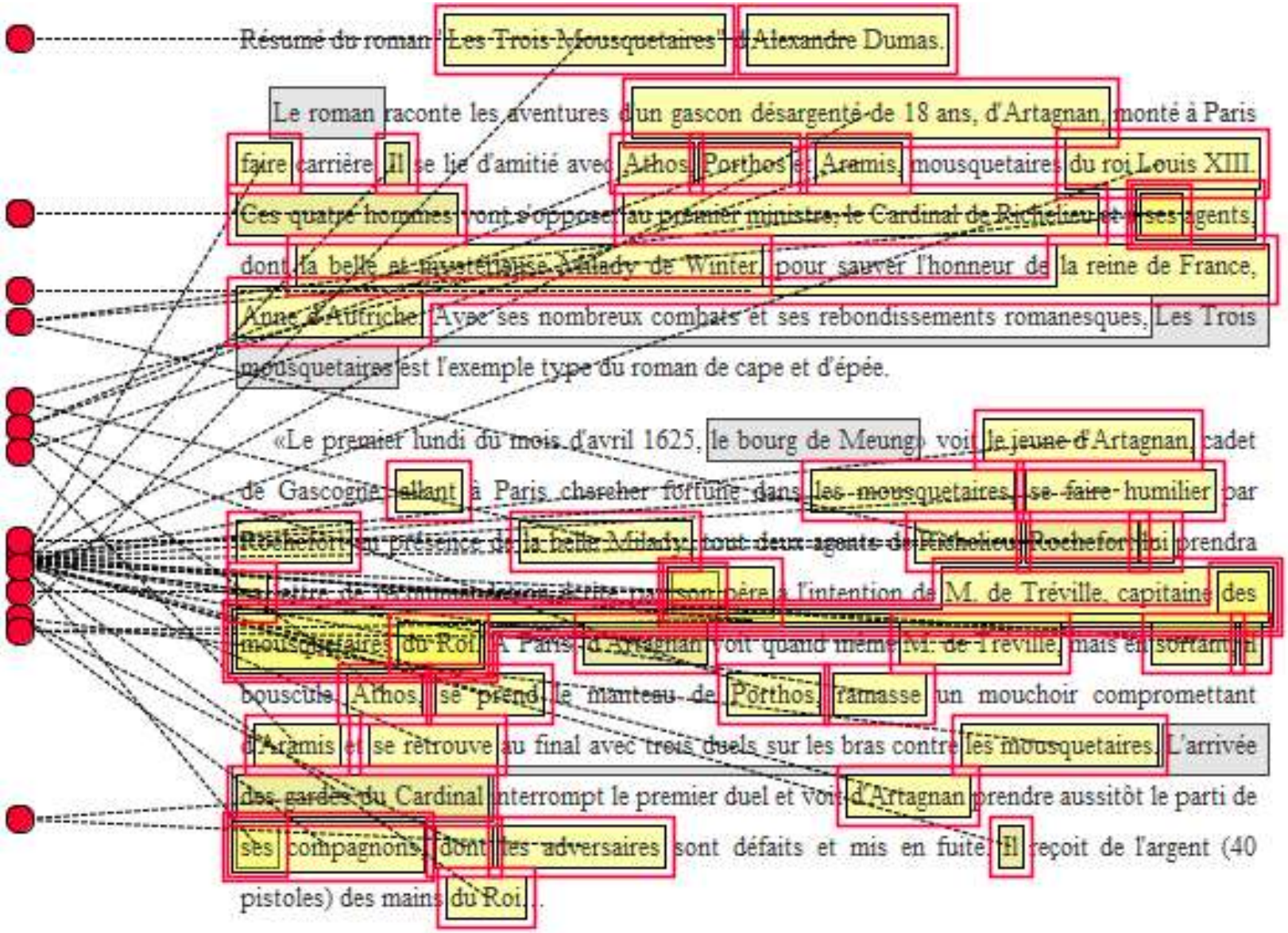
Volet 3 « linguistique outillée » du projet Democrat
http://textometrie.ens-lyon.fr/

Visualisation des chaînes



- Dans la nouvelle *L'occupation des sols* de Jean Echenoz, deux référents très liés :
- Sylvie Fabre
- sa représentation peinte sur un mur

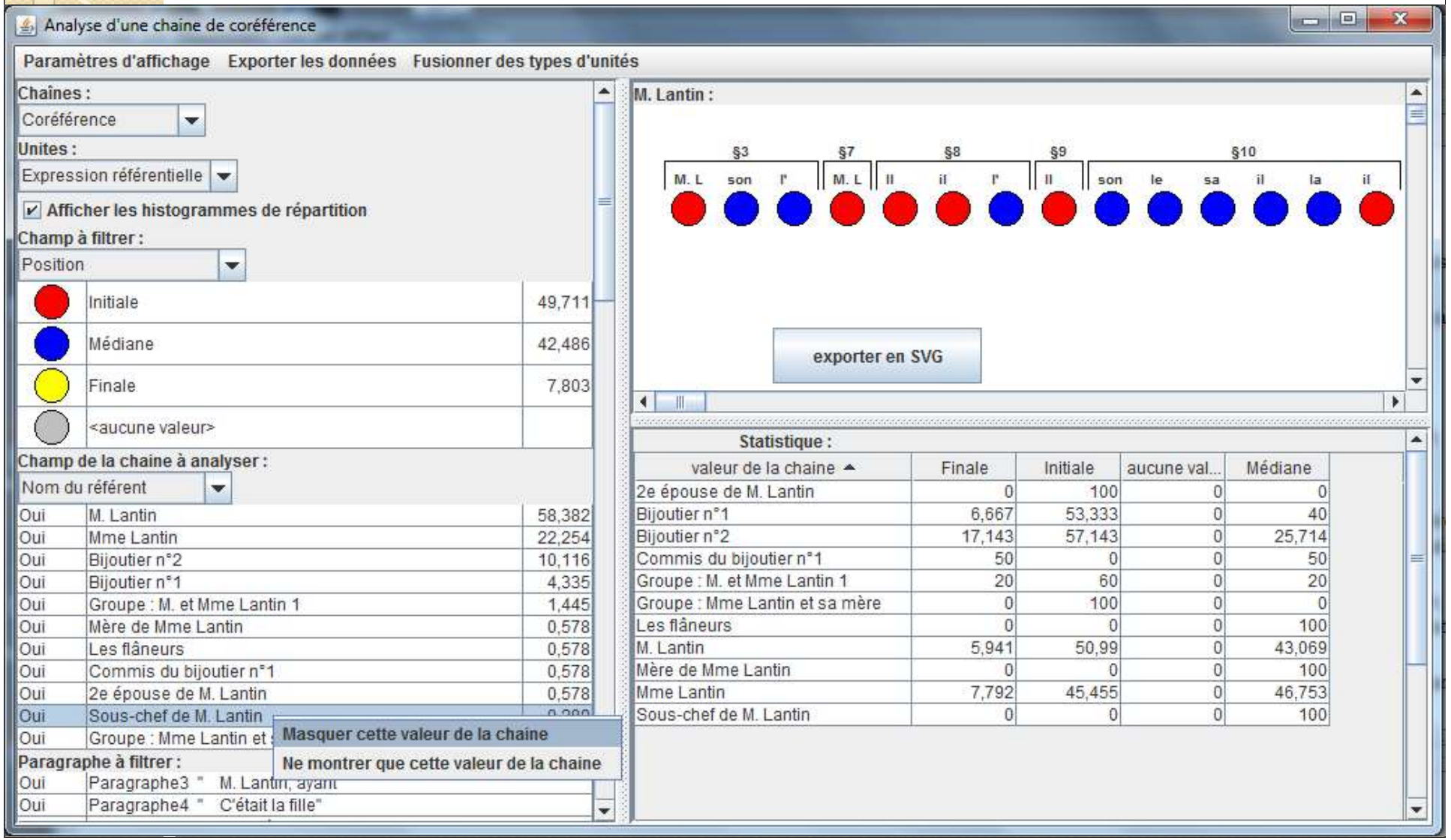
Toutes les chaînes à la fois



Etude manuelle des chaînes

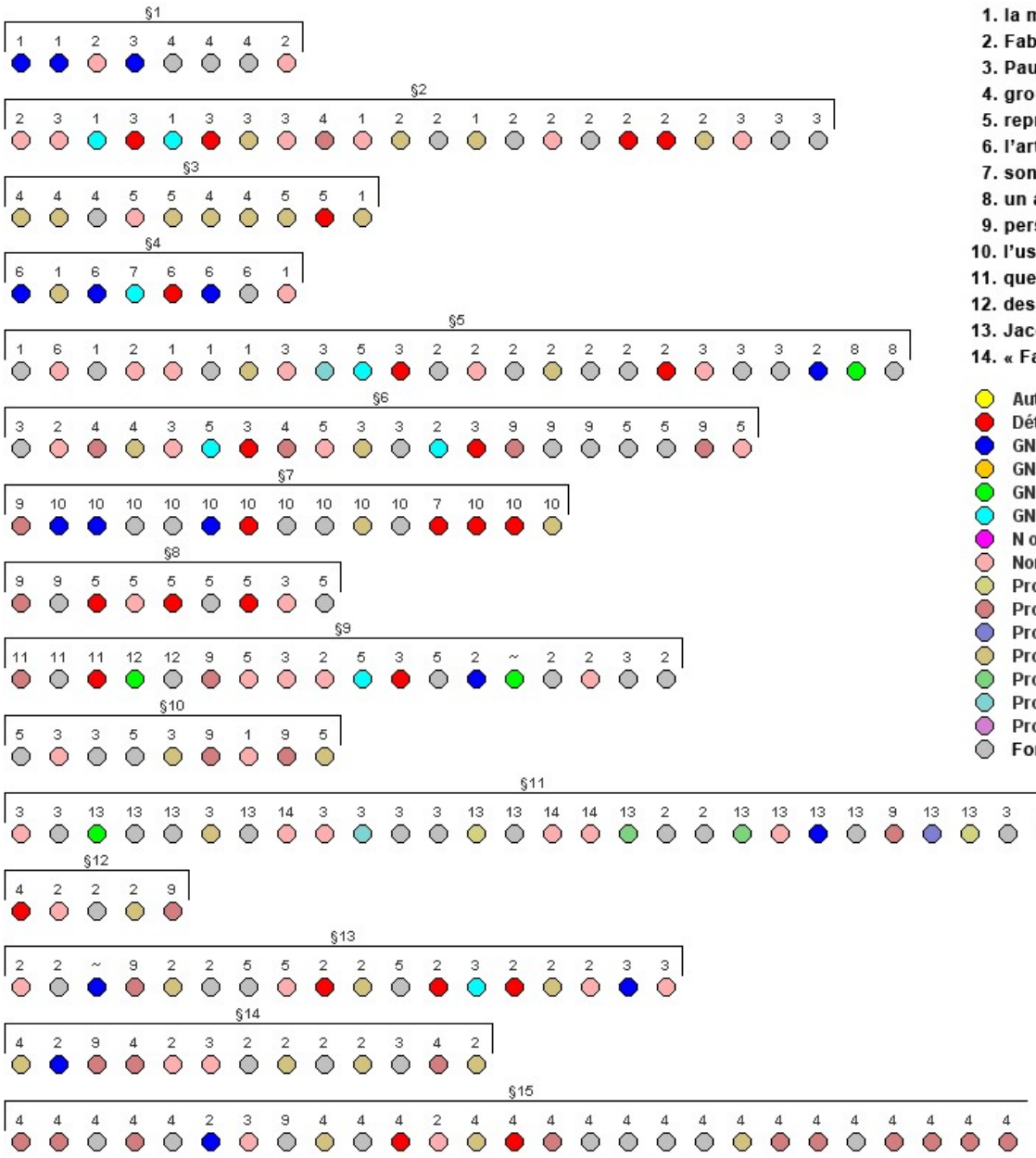
Personnage	Chaîne de coréférence	Proportion de noms propres
Fabre, le père	Fabre – Fabre – Fabre – il – s' – Fabre – sa – ses – l' – Fabre – s' – Fabre – que – il – ses – le père – Fabre – son père – Fabre – le veuf – Fabre – se – il – Fabre – s' – le – ses – il – ses – son – il – Fabre – le père – Fabre – s' – il – se – lui-même – il – le père de Paul – Fabre	32% (13 sur 41)
Paul Fabre, le fils	le fils Paul – Paul – sa – sa – lui – Paul – Paul – Paul – qui – ta – Paul – Paul – sa – il – son – Paul – Paul – sa – se – Paul – il – Paul – lui – Paul – se – Paul – son fils – du fils – il – Paul – Paul – Paul	50% (16 sur 32)
Groupe formé par le père et le fils	se – on – ils – les – eux – on – se – on – leur – on – on – on – on – s' – on – s' – ils – s' – leur – ils – leurs – on – s' – se – on – on – on – on – on – on	0%
La mère	(tout) – la mère – la mère – sa mère – sa mère à lui – Sylvie Fabre – la – elle – l' – Sylvie Fabre – Sylvie – elle – Sylvie	31% (4 sur 13)
L'effigie de la mère	Sylvie Fabre – elle – elle – sa – ta mère – l'effigie – sa mère – Sylvie Fabre – Sylvie – son – Sylvie Fabre – son – ses – Sylvie – sa mère – l' – Sylvie	35% (6 sur 17)
Flers	l'artiste Flers – son – l'artiste Flers – Flers	75%
L'usager	l'usager – l'usager – s' – l'usager – sa – il – se – son – sa – soi	0%
Jacqueline	une femme – qui – s' – celle – qui – j' – tu – Jacqueline – la femme – s' – qui – c'	8%

Etude outillée des chaînes



valeur de la chaine	Finale	Initiale	aucune val...	Médiane
2e épouse de M. Lantin	0	100	0	0
Bijoutier n°1	6,667	53,333	0	40
Bijoutier n°2	17,143	57,143	0	25,714
Commis du bijoutier n°1	50	0	0	50
Groupe : M. et Mme Lantin 1	20	60	0	20
Groupe : Mme Lantin et sa mère	0	100	0	0
Les flâneurs	0	0	0	100
M. Lantin	5,941	50,99	0	43,069
Mère de Mme Lantin	0	0	0	100
Mme Lantin	7,792	45,455	0	46,753
Sous-chef de M. Lantin	0	0	0	100

Etude de la suite des références

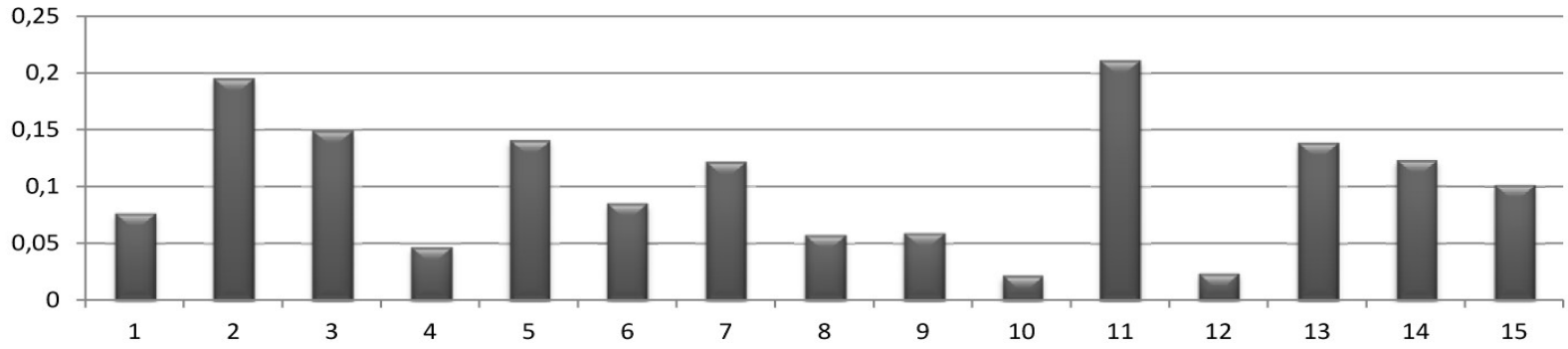


1. la mère
2. Fabre, le père
3. Paul, le fils
4. groupe formé par le père et le fils
5. représentation de la mère
6. l'artiste Flers
7. son équipe
8. un attroupement
9. personne indéfinie
10. l'usager
11. quelqu'un
12. des hommes casqués de jaune
13. Jacqueline
14. « Fabre »

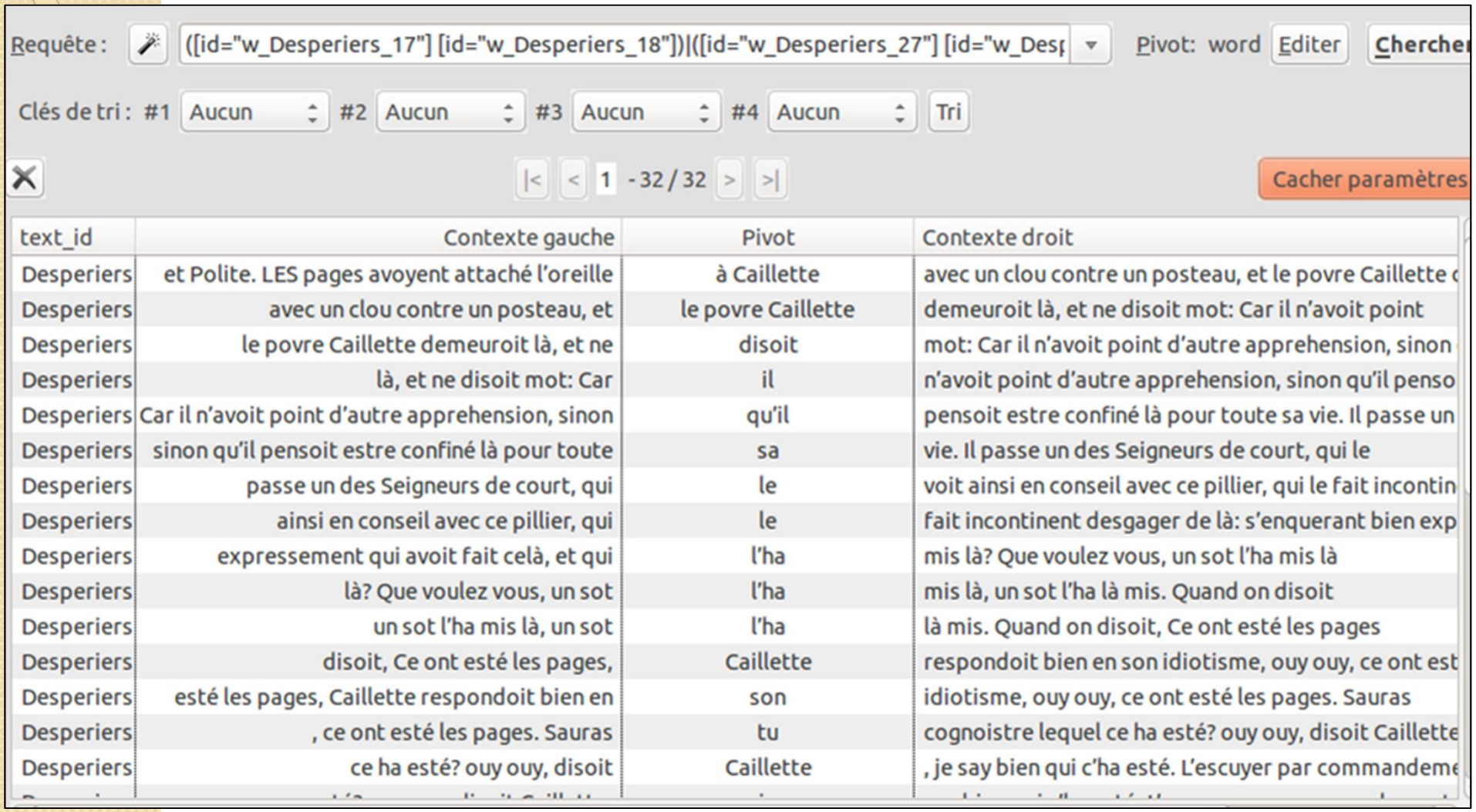
- Autre
- Dét Possessif
- GN Défini
- GN Démonstratif
- GN Indéfini
- GN Possessif
- N ou GN sans dét
- Nom Propre
- Pron Démonstratif
- Pron Indéfini
- Pron Interrogatif
- Pron Pers Anaphorique
- Pron Pers Déictique
- Pron Relatif
- Pron possessif
- Formes atténuées

Etude des densités référentielles

Paragraphe	Titre (en reprenant ceux de l'étude de Catherine Fuchs et Pierre Le Goffic)	Personnages centraux
§1	Incendie et réinstallation du père et du fils	père, fils, mère
§2	Nouvelle vie du père et du fils à l'intérieur	père, fils, mère
§3	Nouvelle vie du père et du fils à l'extérieur	père, fils, effigie
§4	L'immeuble Wagner et l'image de la mère (flashbacks)	Flers
§5	L'image de la mère (flashbacks), retour au père et au fils	père, fils, mère, effigie
§6	Fin de la vie commune du père et du fils ; démolition	père, fils, effigie
§7	Dépérissement de l'espace vert	usager
§8	Dégradation du lieu et de l'image de la mère	effigie
§9	Construction progressive d'un nouvel immeuble	fils
§10	Fin des visites du fils	fils
§11	Retrouvailles du fils avec le père, installé…	Jacqueline
§12	Dans le nouvel appartement du père	père
§13	Flashback sur l'emménagement du père	père
§14	Retour du fils pour le week-end	père, fils
§15	Déjeuner, puis attaque du grattage	père et fils

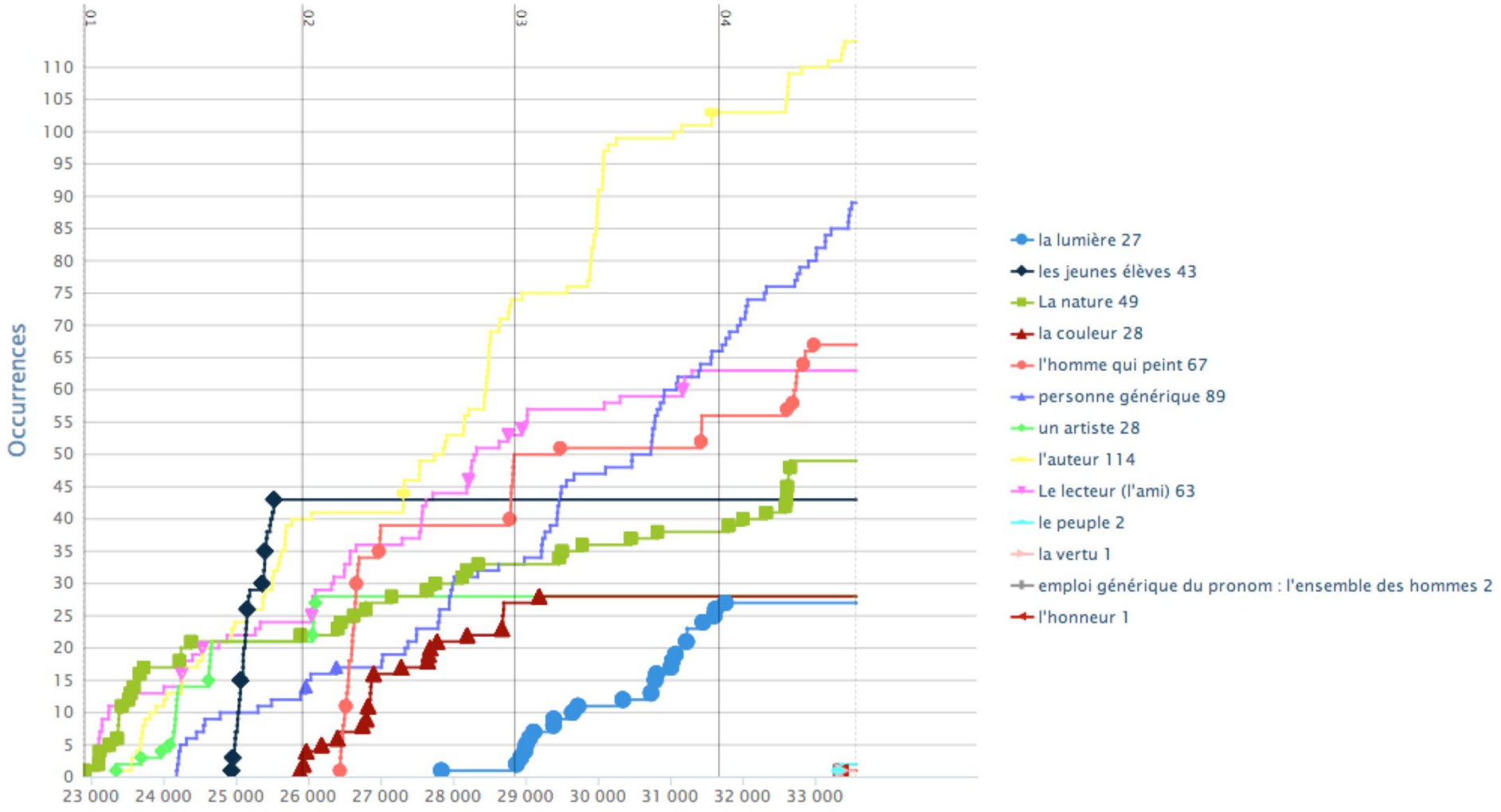


Concordancier appliqué aux chaînes



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

Diagramme de progression



Bilan

- Evolutions de TXM liées à Democrat
  - Fonctionnalités d'annotation fondées sur le modèle URS
  - Nouveau format de fichier XML TEI compatible avec le modèle URS
  - Interconnexion de ces fonctionnalités avec les autres modules de TXM (ce qui a permis par exemple de mettre en œuvre le concordancier)
  - Fonctionnalités de visualisation des chaînes de référence
  - Publication de l'extension « URS » (Democrat) dans TXM 0.8.0
- Autres outils
  - Interface d'interrogation de chaînes d'annotations dans ANALEC
  - SACR : script d'annotation de chaînes de référence (outil d'annotation en ligne développé en javascript)
  - CRViewer : outil de visualisation et de calcul de statistiques descriptives simples sur les chaînes de référence
  - Publications : LREC 2018, JADT 2016 et JADT 2018

Traitement automatique des langues : détection automatique des chaînes de coréférences

Volet 4 « TAL » du projet Democrat
https://github.com/boberle/cofr
https://github.com/LoicGrobol

Voie 1 : systèmes de règles

- Systèmes à base de règles
  - Principe : on écrit à la main un ensemble de règles :
    - Si article défini alors…
    - Si distance entre deux mentions inférieure à 4 mots, alors…
  - Avantage : les règles sont lisibles (compréhensibles) et sont issues d'une collaboration entre linguistes et informaticiens
  - Inconvénients :
    - Manquent de souplesse : toute correction de règle peut avoir des effets collatéraux et dégrader les performances globales
    - Manquent de performance, surtout pour des tâches complexes, impliquant de nombreux paramètres

- Note au passage (phase 2 de la procédure d'annotation)
  - C'est un système de règles qui est utilisé pour annoter automatiquement la catégorie des expressions référentielles

Voie 2 : apprentissage artificiel

- On confie à un système :
  - La détermination de ses propres règles
  - La détermination de ses propres seuils (distance entre 2 mentions)
  - Avantages : grande souplesse, peu d'intervention de l'intuition seule
  - Inconvénients :
    - Les solutions trouvées par le système sont parfois peu lisibles et non modifiables a posteriori : on doit les accepter telles quelles
    - Les approches hybrides (règles + apprentissage) sont difficiles à mettre en œuvre : il vaut mieux relancer un apprentissage…
    - Surtout : le système apprend à partir d'une base : un corpus annoté qui fait référence (on ne peut pas apprendre sans corpus annoté)

- Sur les chaînes de référence en français
  - Corpus ANCOR → système CROC ; systèmes CRAC / DeCOFRe
  - Corpus DEMOCRAT → système COFR

Apprentissage et application

- Phase d'apprentissage : corpus annoté → modèle
  - On prend un corpus annoté
  - On le scinde en plusieurs parties : une pour l'apprentissage proprement dit, les autres pour test et validation
  - Des annotations, on extrait des exemples avec leurs caractéristiques
  - Le **système d'apprentissage** apprend à partir de ces exemples
  - Pour éviter le sur-apprentissage, on force le modèle à ne pas être « trop proche » des données, pour encourager la généralisation (régularisation)

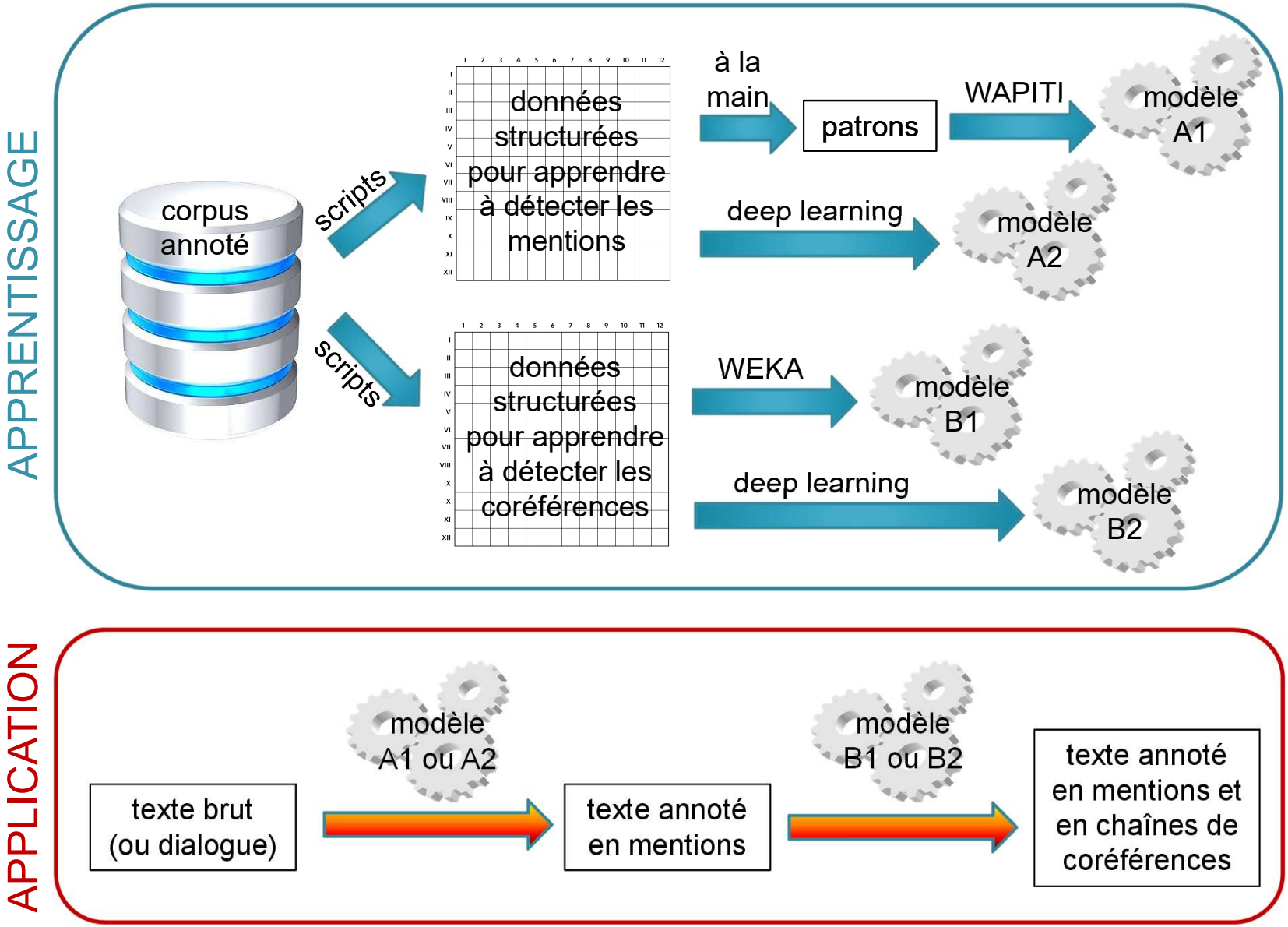
- Phase d'application : texte brut $\xrightarrow{\text{modèle}}$ texte annoté
  - On prend un texte brut
  - On lui applique le modèle d'apprentissage
  - On obtient un texte annoté automatiquement
  - C'est le **système *end-to-end***

« Nourrissage » du système d'apprentissage des coréférences

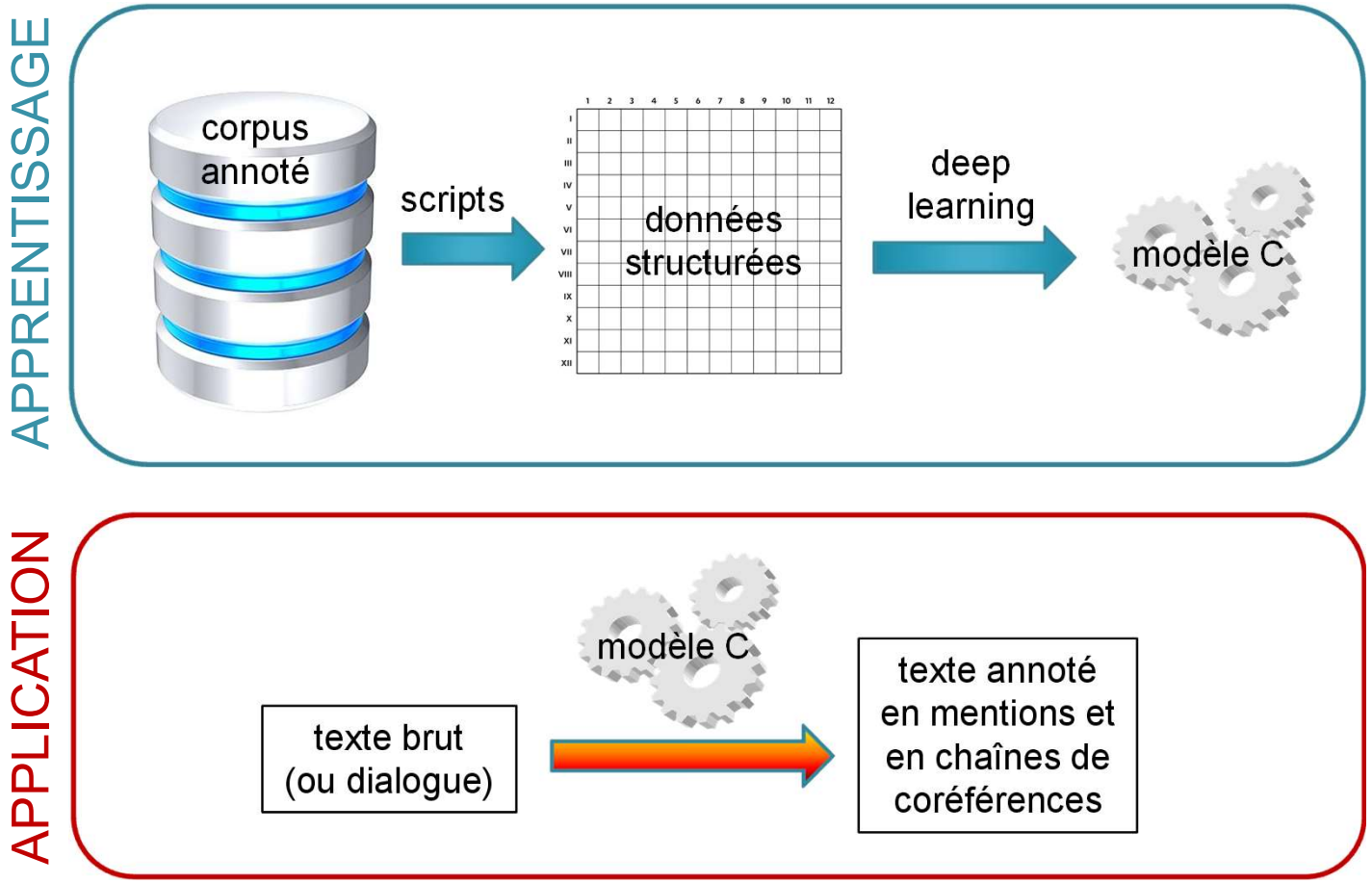
- Principe général
  - On identifie des paramètres qui pourraient aider l'apprentissage
  - On calcule des traits (*feature*)
  - On fournit un fichier (potentiellement énorme) au système

- Tout se fait avec des traits
  - On peut en imaginer autant qu'on veut, mais encore faut-il les calculer, notamment pour le système *end-to-end*…
  - A terme, on pourrait envisager un système hybride avec à la fois de l'apprentissage et des règles, en amont ou en aval (mais concilier les deux est loin d'être immédiat)

Approche « scindée »



Approche « globale »



Résultats : 75% avec CoNLL

[**Le marquis**]$_1$ professait [une haine vigoureuse] pour [les lumières] ; ce sont [**les idées**]$_2$, disait [**-il**]$_1$, [**qui**]$_2$ ont perdu [l' Italie] ; [**il**]$_1$ ne savait trop comment concilier [cette sainte horreur de [l' instruction]] , avec le désir de voir [[**son**]$_1$ **fils Fabrice**]$_3$ perfectionner [l' éducation si brillamment commencée chez [les jésuites]] . Pour courir [le moins de risques possible] , [**il**]$_1$ chargea [**le bon abbé Blanès**]$_4$, curé de [Grianta] , de faire continuer [**Fabrice**]$_3$ [**ses**]$_3$ études en [**latin**]$_5$. Il eût fallu que [**le curé lui-même**]$_4$ sût [**cette langue**]$_6$; or [**elle**]$_6$ était [l' objet de [[**ses**]$_3$ mépris]] ; [[**ses**]$_3$ connaissances en [ce genre]] se bornaient à réciter , par cœur , [les prières de [[**son**]$_3$ **missel**]$_7$] , [**dont**]$_7$ [**il**]$_3$ pouvait rendre à peu près [le sens] à [[**ses**]$_3$ ouailles] . Mais [**ce curé**]$_4$ n' en était pas moins fort respecté et même redouté dans [le canton] ; [**il**]$_4$ avait toujours dit que ce n' était point en [treize semaines] ni même en [treize mois] , que l' on verrait s' accomplir [la célèbre prophétie de [saint Giovita]] , le patron de [Brescia] . [**Il**]$_4$ ajoutait , quand [**il**]$_4$ parlait à [des amis sûrs] , que [ce nombre treize] devait être interprété d' [**une façon**]$_8$ [**qui**]$_8$ étonnerait bien de [**le monde**]$_9$, s' il était permis de tout dire ([1813]) . [Le fait] est que [**l' abbé Blanès**]$_4$, personnage d' [[une honnêteté] et d' [une vertu primitives]] , et de plus homme d' esprit , passait [toutes les nuits] à [le haut de [[**son**]$_4$ **clocher**]$_{10}$] ; [**il**]$_4$ était fou d' [astrologie] . Après avoir usé [[**ses**]$_4$ journées] à calculer [[**des conjonctions**] **et** [**des positions d' étoiles**]]$_{11}$, [**il**]$_4$ employait [la meilleure part de [[**ses**]$_4$ nuits]] à [**les**]$_{11}$ suivre dans [le ciel] . Par suite de [[**sa**]$_4$ pauvreté] , [**il**]$_4$ n' avait d' [autre instrument] qu' [une longue lunette à [tuyau de carton]] . [**On**]$_{12}$ peut juger de [**le mépris**]$_{13}$ [**qu'**]$_{13}$ avait pour [l' étude de [les langues]] [**un homme**]$_{14}$ [**qui**]$_{14}$ passait [[**sa**]$_{14}$ vie] à découvrir [l' époque précise de [**la chute de** [**les empires**] **et de** [**les révolutions**]]$_{15}$] [**qui**]$_{15}$ changent [la face de [**le monde**]$_9$] . Que sais [**-je**]$_{16}$ de plus sur [**un cheval**]$_{17}$, disait [**-il**]$_{16}$ à [**Fabrice**]$_3$, depuis qu' [**on**]$_{12}$ [**m'**]$_{16}$ a appris qu' en [**latin**]$_5$ [**il**]$_{17}$ s' appelle [equus] ?

Bilan et perspectives

Bilan

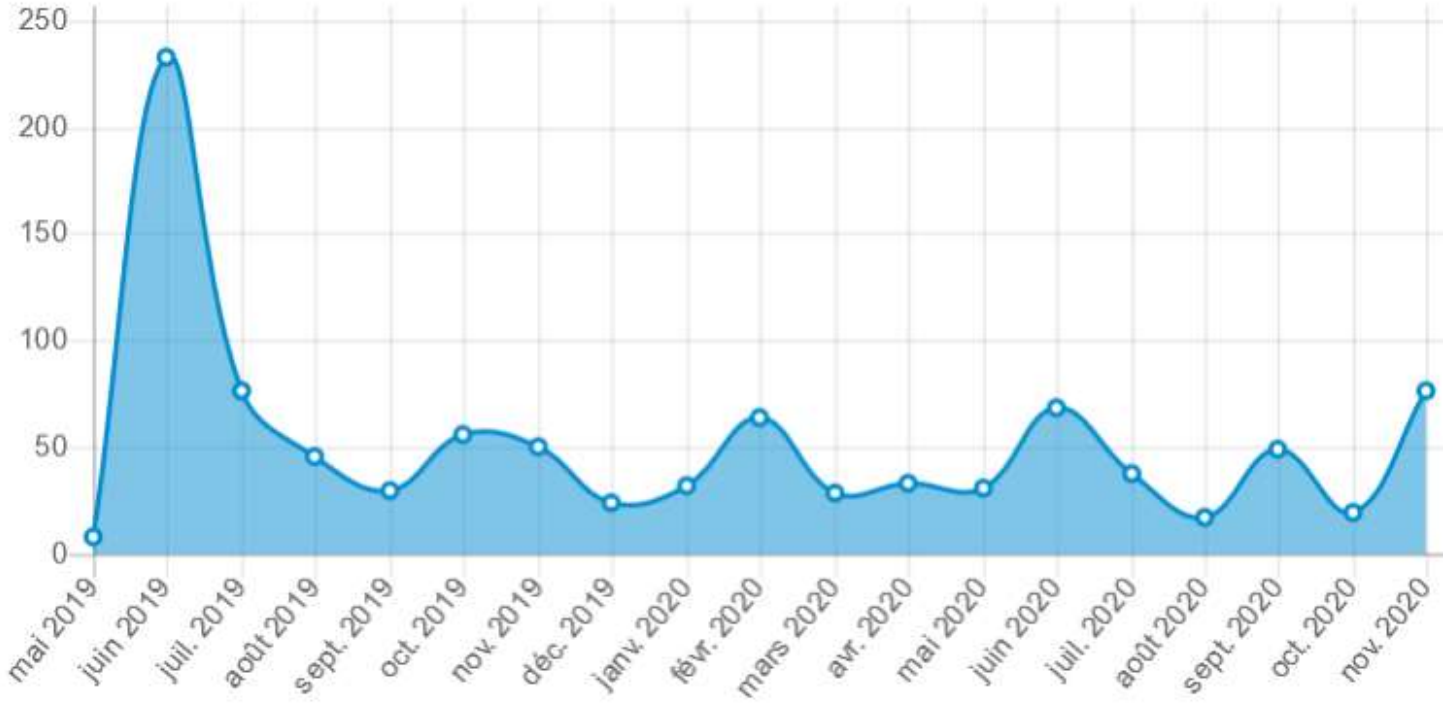
- Mise à disposition à une large communauté de données enrichies et de nouvelles connaissances sur la langue : FAIT
- Mise à disposition de nouveaux outils et de nouveaux procédés de visualisation pour la manipulation de ces données et connaissances : FAIT
- Mise à disposition de nouvelles méthodes d'analyse (linguistique et statistique) des chaînes de coréférences : FAIT
- Représentation de systèmes de TAL traitant la langue française dans des campagnes d'évaluation internationales : pas encore…
- Contribution aux humanités numériques :
  - à la pérennisation des données, à leurs standardisation : FAIT
  - à la place du français dans le monde : ?
  - à la didactique et à l'enseignement du français et des langues : ?

Publications Democrat

- Répartition des publications par objectif
  - Objectif 1 = 9 articles, de nombreux soumis, plusieurs en préparation
  - Objectif 2 = 6 articles, plusieurs soumissions envisagées
  - Objectif 3 = 5 articles + 5 logiciels, quelques soumissions envisagées
  - Objectif 4 = 11 articles, plusieurs soumissions envisagées
  - Dissémination = 3 articles, 3 posters
- Remarques
  - L'objectif 1 (modélisation) était le plus faible au cours du projet, c'est maintenant celui pour lequel il y a le plus d'évolutions à venir
  - L'objectif 4 (TAL) comprend à la fois des articles sur la détection des chaînes et des articles de recherche sur les architectures de réseaux neuronaux artificiels (hors application aux CR), ce qui nous a valu une remarque lors de l'évaluation intermédiaire du projet (hors périmètre)
  - On est fiers d'avoir publié sur des aspects connexes !

Indicateurs : corpus

Statistiques de consultation (972 vues au total)



Statistiques de téléchargements (depuis novembre 2016 ℹ)

97 téléchargements complets
7 fichiers ou dossiers téléchargés

Indicateurs : h-index$_{Democrat}$ = 7

Le volet TAL est le plus cité (et l'a toujours été), mais pas seulement

	Cites	Per year	Rank	Authors	Title	Year	Publication
☑ h	15	5.00	20	Y Dupont, M Dinarelli, I Tellier	Label-dependencies aware recurrent neural networks	2017	International Con...
☑ h	12	4.00	6	M Dinarelli, V Vukotić, C Raymond	Label-dependency coding in simple recurrent networks...	2017	
☑ h	11	3.67	22	C Schnedecker, J Glikman, F Landragin	Les chaînes de référence: annotation, application et qu...	2017	Langue française
☑ h	11	3.67	28	C Schnedecker	Les chaînes de référence: une configuration d'indices p...	2017	Langue française
☑ h	9	2.25	11	A Désoyer, F Landragin, I Tellier, A Lef...	Coreference resolution for french oral data: Machine le...	2016	... on Intelligent T...
☑ h	9	2.25	25	F Landragin	Conception d'un outil de visualisation et d'exploration ...	2016	
☑ h	7	7.00	9	L Grobol	Neural coreference resolution with limited lexical conte...	2019	
☑	7	3.50	12	B Oberle	SACR: A drag-and-drop based tool for coreference ann...	2018	Proceedings of th...
☑	5	1.67	40	V Obry, J Glikman, C Guillot-Barbance,...	Les chaînes de référence dans les récits brefs en françai...	2017	Langue française
☑	4	2.00	7	L Grobol, I Tellier, ÉV De La Clergerie, ...	ANCOR-AS: Enriching the ANCOR corpus with syntactic ...	2018	LREC 2018-11th ...
☑	4	1.33	15	F Landragin, J Potier, M Bothua	Annotation manuelle d'expressions référentielles: expér...	2017	
☑	4	4.00	17	M Dinarelli, L Grobol	Seq2biseq: Bidirectional output-wise recurrent neural n...	2019	arXiv preprint arX...
☑	4	4.00	32	C Schnedecker	De l'intérêt de la notion de chaîne de référence par rap...	2019	Cahiers de praxé...
☑	3	1.00	10	L Grobol, F Landragin, S Heiden	Interoperable annotation of (co) references in the Dem...	2017	
☑	3	3.00	14	M Dinarelli, L Grobol	Hybrid neural models for sequence modelling: The bes...	2019	arXiv preprint arX...
☑	3	1.00	26	L Grobol, I Tellier, ÉV de La Clergerie, ...	Apports des analyses syntaxiques pour la détection aut...	2017	
☑	3	1.00	36	E Baumer	Chaînes de référence et point de vue dans la fiction litt...	2017	Langue française
☑	3	1.00	37	F Landragin	Analyse, visualisation et identification automatique des...	2017	Langue française
☑	2	1.00	1	S Heiden	Annotation-based Digital Text Corpora Analysis within ...	2018	КОРПУСНАЯ ЛИ...
☑	2	0.67	4	B Oberle	Coreference annotation with SACR, a new drag-and-dr...	2017	ECLAVIT Workshop
☑	2	2.00	24	R Wilkens, B Oberle, F Landragin, ...	French coreference for spoken and written language	2020	Language Resour...
☑	2	1.00	41	F Landragin	Étude de la référence et de la coréférence: rôle des pet...	2018	Corpus

Perspectives

- 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