ANR-15-CE38-0008

DEscription et MOdélisation des Chaînes de Référence : outils pour l'**A**nnotation de corpus (en diachronie et en langues comparées) et le **T**raitement automatique

Cultures, patrimoines, création (DS0805) – Edition 2015

LIVRABLE L2

« Manuel d'annotation du corpus et organisation de formations sur l'annotation »

mars 2018

Editeur : Frédéric Landragin (Lattice, coordinateur du projet)

Contributeurs principaux : Meryl Bothua (Lattice), Frédéric Landragin (Lattice), Bruno Oberlé (LiLPa), Céline Guillot-Barbance (IHRIM), Catherine Schnedecker (LiLPa), Matthieu Quignard (ICAR), Zeina Tmart (ENS Lyon).

Sommaire

A IDENTIFICATION DU PROJET

Acronyme du projet	DEMOCRAT
Titre du projet	DEscription et MOdélisation des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique
Coordinateur du projet (société/organisme)	Frédéric Landragin (Lattice-ENS-CNRS-Université Sorbonne Nouvelle – Paris 3)
Date de début du projet Date de fin du projet	1er mars 2016 (*T0 scientifique*) 29 février 2020 (*Tfinal scientifique*)
Labels et correspondants des pôles de compétitivité (pôle, nom et courriel du corresp.)	–
Site web du projet, le cas échéant	http://www.lattice.cnrs.fr/democrat/

Rédacteur de ce rapport	
Civilité, prénom, nom	Frédéric Landragin
Téléphone	01 58 07 66 20
Courriel	frederic.landragin@ens.fr
Date de rédaction	Février 2018
Période faisant l'objet du rapport d'activité	Du 1er mars 2016 au 28 février 2018

B CONTENU DU DOCUMENT

Ce document est chronologiquement le premier livrable (bien que nommé « L2 ») du projet Democrat : il concerne la méthodologie d'annotation manuelle du corpus Democrat, et sert de document préalable au livrable L1 : « Mise à disposition du corpus annoté sur le site web du projet (et information via les listes de diffusion de la communauté linguistique et traitement automatique des langues) » – à rendre en mars 2019.

Ce document n'est pas voué à être public. En revanche, le livrable L1 le sera, et sera accompagné du manuel d'annotation du corpus, c'est-à-dire de l'ensemble de la partie D du présent document. Il contiendra également quelques mesures d'accord inter-annotateurs, qui n'apparaissent pas ici dans la mesure où l'annotation du corpus n'est pas encore finalisée.

Ce document est divisé en 5 sections :

- A : identification du projet (comme pour les rapports d'avancement intermédiaires),
- B : synthèse rapide du contenu du document et de ses objectifs,

- C : méthodologie générale du projet, qui – comme écrit plus haut – n'est pas vouée à être diffusée publiquement, mais a en revanche déjà fait l'objet de publications (déposées sur HAL[1] et donc diffusées publiquement),
- D : manuel d'annotation, qui sera joint au livrable L1,
- E : récapitulatif des formations déjà effectuées, formations qui concerne à la fois la bonne exploitation du manuel d'annotation (D) mais aussi l'utilisation des outils d'annotation, notamment TXM[2] qui fera l'objet du livrable L3a (à rendre en mars 2019).

C METHODOLOGIE GENERALE DU PROJET

C.1 METHODOLOGIE DE CHOIX DES PHENOMENES LINGUISTIQUES

Le projet Democrat porte sur la référence et les chaînes de référence. La référence est un objet linguistique très vaste, qui a fait l'objet de très nombreuses publications[3]. Le point de départ est la notion d'« expression référentielle », c'est-à-dire d'un mot ou suite de mots (comme « le président de la république ») qui renvoie à un objet issu du monde réel ou imaginaire.

Dans son étude des expressions référentielles, le projet MC4[4] s'était intéressé aux multiples facteurs morphologiques, syntaxiques, sémantiques et pragmatiques qui interviennent lors de la résolution des références, c'est-à-dire l'attribution d'un référent à une expression référentielle, en incluant l'attribution d'un antécédent à une anaphore.

---

[1] Landragin, F. (2017) Analyse, visualisation et identification automatique des chaînes de coréférences : des questions interdépendantes ?, *Langue Française* 195, pp. 17-34 (18 pages) [https://halshs.archives-ouvertes.fr/halshs-01580784]. Schnedecker, C., Glikman, J. & Landragin, F. (2017) Les chaînes de référence : annotation, application et questions théoriques, *Langue Française* 195, pp. 5-15 (11 pages) [https://halshs.archives-ouvertes.fr/halshs-01580785]. Landragin, F. (2017), Potier, J. & Bothua, M., Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus, In: *Neuvièmes Journées Internationales de la Linguistique de Corpus (JLC 2017)*, Grenoble, pp. 43-46 (4 pages) [https://halshs.archives-ouvertes.fr/halshs-01513810]. Landragin F. (2016) Conception d'un outil de visualisation et d'exploration de chaînes de coréférences, In: *Thirteen International Conference on Statistical Analysis of Textual Data (JADT 2016)*, Nice, pp. 109-120 [https://halshs.archives-ouvertes.fr/halshs-01329414]. Grobol, L., Landragin, F. & Heiden, S. (2017) Interoperable annotation of (co)references in the Democrat project, In: *Thirteenth Joint ISO–ACL Workshop on Interoperable Semantic Annotation (ISA'13), Workshop Section at the twelfth International Conference on Computational Semantics (IWCS)*, Montpellier, pp. 99-105 (7 pages) [https://hal.archives-ouvertes.fr/hal-01583527v2].

[2] Heiden S. & Decorde M. (2017) TXM Software, logiciel [https://halshs.archives-ouvertes.fr/halshs-00377694].

[3] Citons entre autres la synthèse suivante : Charolles, M. (2002) *La Référence et les expressions référentielles en français.* Paris : Ophrys.

[4] Projet d'une durée de deux ans qui a été un préalable au projet Democrat, cf. notamment – pour la méthodologie d'annotation – l'article : Landragin, F. (2011) Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits, *Corpus* 10, http://journals.openedition.org/corpus/, pp. 61-80.

La procédure d'annotation résultante avait impliqué l'annotation manuelle de ces facteurs, du moins d'une sélection d'une dizaine de facteurs considérés comme déterminants. Au final, le corpus n'a pas dépassé 5 000 expressions référentielles annotées.

Pour le projet Democrat, il n'était pas question de reproduire une procédure aussi détaillée, et ce d'autant plus que l'annotation de certains de ces facteurs est automatisable. Nous avons ainsi choisi d'annoter uniquement le résultat de la résolution de la référence. Deux possibilités apparaissent ici :

- soit on saisit, pour chaque expression, un identifiant du référent,
- soit on regroupe les expressions en chaînes (anaphoriques et/ou coréférentielles), ce qui fait l'économie des identifiants des référents mais nécessite de construire des chaînes, autrement dit des objets non liés à un et un seul marquable.

Une première expérimentation a permis de comparer les deux méthodes et a soulevé l'importance décisive de l'ergonomie de l'outil d'annotation utilisé : comme il est possible de déduire automatiquement les chaînes à partir d'une annotation des mentions en identifiants, seule compte la rapidité d'action. Or, quand l'outil est bien choisi et permet la complétion automatique de l'identifiant en cours de saisie, il s'avère que la méthode à base d'identifiants est plus rapide que celle à base de construction de chaînes. En effet, manipuler un objet couvrant potentiellement le texte entier est bien plus délicat et propice à des erreurs que saisir des identifiants localement, au niveau du marquable qu'est l'expression référentielle.

Plusieurs outils ont été testés (MMAX2[5], GLOZZ[6], ANALEC[7]) et c'est finalement l'implémentation de la complétion dans l'outil ANALEC qui a permis la plus grande efficacité. C'est pourquoi le projet Democrat a décidé de retenir ANALEC dans la mise en œuvre d'un chantier d'évolution de la plateforme TXM.

Pour conclure, le choix des phénomènes linguistiques est le suivant :

1. **Les expressions référentielles, directement *via* les marquables**, c'est-à-dire les traces linguistiques d'un accès à un référent présent dans le monde extra-linguistique, que ce référent soit un personnage humain, un objet concret ou abstrait.
2. **Les chaînes de référence, de manière indirecte** : comme nous le verrons dans la section suivante, chaque expression référentielle est annotée avec un identifiant de référent, et c'est cette information qui permet de construire (puis d'étudier) les chaînes de référence.

[5] Müller, C., Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J. (Eds.). *Corpus technology and language pedagogy: New resources, new tools, new methods*. Frankfurt : Peter Lang.

[6] Widlöcher, A., Mathet, Y. (2012). The Glozz platform: A corpus annotation and mining tool. In: *Proceedings of the 2012 ACM Symposium on Document Engineering*, ACM, 171-180.

[7] Landragin, F., Poibeau, T., Victorri, B. (2012). ANALEC: a New Tool for the Dynamic Annotation of Textual Data. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turquey, 357-362.

C.2 METHODOLOGIE DE CHOIX DES ANNOTATIONS

Une annotation est une structure de « traits » (couples attributs-valeurs) que l'on affecte aux phénomènes étudiés, soit 1. les expressions référentielles ; 2. les chaînes de référence.

Contrairement au projet MC4 dont nous avons parlé et qui a suivi une méthodologie consistant à affecter de nombreux traits aux expressions référentielles, **nous avons décidé dans le projet Democrat de n'affecter qu'un seul trait : l'identifiant du référent** (sous le nom de « REF »). C'est un choix fort mais motivé par plusieurs raisons :

1. Moins l'annotateur a de champs à remplir, plus rapide sera la tâche d'annotation. Or le corpus visé pour Democrat est un corpus de grande taille, dont l'annotation manuelle doit être limitée temporellement.

2. Les champs présents dans le projet MC4 relevaient d'aspects morphosyntaxiques et syntaxiques. Or il nous a semblé que beaucoup de ces aspects sont automatisables et ne doivent donc pas relever d'une annotation manuelle. Notamment quand l'expression référentielle est déjà délimitée par l'annotateur : on est en présence de formes de surface telles que « le chat », pour lesquelles on peut imaginer le développement de macros afin d'ajouter un champ « type d'expression » qui contienne la valeur « groupe nominal défini » ou encore les champs genre et nombre avec masculin et singulier. Autrement dit : rien qu'en délimitant les expressions référentielles, l'annotateur facilite le développement de macros permettant l'ajout d'annotations automatiques fiables.

Par ailleurs, n'avoir qu'un champ à remplir permet de focaliser la tâche sur le repérage et la délimitation des expressions référentielles. Et c'est sur ce point qu'ont eu lieu de nombreuses discussions ainsi que des choix parfois difficiles à faire.

Ainsi, nos premières expérimentations d'annotation se sont déroulées sur des textes narratifs, en l'occurrence des extraits de romans libres et gratuits, disponibles sur la plateforme « wikisource ». Il s'agit donc de textes littéraires, et les discussions ont vite porté sur le filtrage ou non des référents : faut-il annoter toutes les expressions référentielles ? ou seulement celles qui réfèrent à des personnages humains ? à des êtres animés ? à des objets concrets ? Nous constatons par exemple que les expressions référentielles temporelles sont très peu reprises et ne forment donc que rarement des chaînes de référence intéressantes à étudier. Dans ce cas, faire l'impasse sur l'annotation de ces « singletons » permettrait d'aller plus vite à l'essentiel.

Sauf que :

1. l'annotation systématique de toutes les expressions référentielles permet de nourrir un système d'apprentissage dédié à la détection des expressions référentielles (ce qui, en TAL, est une tâche très complexe, différente de celles consistant à détecter les entités nommées et les pronoms anaphoriques) ;

2. il n’est pas possible de savoir à coup sûr si l’expression en cours d’annotation va être reprise ultérieurement ou non (autrement dit la tâche peut comporter des retours en arrière dans le texte si l’annotateur s’aperçoit qu’il a oublié une expression – initialement considérée comme singleton) ;
3. se demander à chaque expression si elle a des chances d’être un singleton ou de faire partie d’une chaîne de référence va à l’encontre de l’aspect « robotique » et efficace de l’annotation : se poser trop de questions est parfois contre-productif, et il vaut mieux tout annoter en se posant moins de questions…

Au final, nous avons convenu de rendre le choix des identifiants de référents le plus rapide possible pour les expressions peu susceptibles d’être reprises. Pour les expressions qui resteront clairement des singletons, nous employons un code dédié (« SI », comme singleton), ce qui permet d’augmenter encore l’efficacité.

En revanche, dans la mesure où le corpus Democrat a trois exploitations majeures visées – premièrement en tant que réservoir d’exemples attestés pour des analyses linguistiques variées, deuxièmement en tant que données permettant des analyses quantitatives (statistiques voire textométriques), et troisièmement en tant que corpus d’apprentissage pour des visées de traitement automatique des langues – **nous avons décidé d’annoter toutes les expressions référentielles, sans faire de filtrage**.

Concernant les chaînes de référence, il faut tout d’abord noter qu’elles sont théoriquement bien moins nombreuses que les expressions référentielles. Il est donc *a priori* non nécessaire de limiter absolument le nombre de traits de l’annotation des chaînes. C’est pourquoi, même si c’est un point qui n’a pas encore été validé expérimentalement, **nous avons décidé d’annoter toutes les chaînes avec un ensemble de traits caractérisant le référent concerné**.

C.3 Experimentations chronometrees pour valider la methodologie

Afin de valider la faisabilité de cette décision, nous avons malgré tout réalisé quelques expérimentations impliquant une sélection des expressions référentielles à annoter. La sélection a dans un premier temps reposé sur la nature des référents – humains et animaux, par exemple – et dans un deuxième temps sur les thèmes explorés dans le texte : une expérimentation avec des extraits de presse écrite (*L’Est Républicain*) et avec des fiches issues de l’encyclopédie Wikipédia a montré que le choix des référents intéressants revenait surtout à l’annotateur lors de sa lecture, et non à une catégorisation *a priori*. Nous avons un moment envisagé d’annoter en deux étapes : une première pour les référents pertinents, et – si besoin – une seconde pour l’ensemble des référents. Cette procédure s’est révélée peu convaincante, notamment parce que certains annotateurs ressentent comme difficile la reprise de leurs annotations en vue de les augmenter : il est préférable pour certains d’accomplir la totalité de la tâche en un seul passage.

A titre d'exemple, le tableau 1 récapitule les résultats de sept expérimentations chronométrées (30 minutes par annotateur et par texte), en suivant différentes stratégies. Dans les colonnes 2 et 3, le premier chiffre est le nombre de mentions repérées et annotées, et le deuxième est le rang du mot atteint dans le texte (au bout de 30 minutes, donc). Cela donne un rythme moyen de 500 mentions annotées dans une journée pleine, sachant que cette expérimentation a été réalisée après plusieurs jours d'annotations, de manière à ce que les annotateurs aient bien en tête les subtilités du schéma d'annotation. Autrement dit, le rythme donné est un rythme de « croisière » et ne reflète pas du tout la vitesse d'un annotateur débutant.

Textes	Annotateur 1	Annotateur 2	Stratégie suivie
Le collier des jours	95 – 386 mots	93 – 541 mots	Systématique
Boule de suif	100 – 392 mots	78 – 348 mots	Systématique
L'enfant	114 – 338 mots	111 – 345 mots	Systématique
La recherche de l'absolu	85 – 275 mots	80 – 301 mots	Systématique
Manon Lescaut	120 – 397 mots	100 – 390 mots	Objets, humains et animaux
Douce lumière	105 – 543 mots	81 – 311 mots	Systématique avec chunks
Le Capitaine Fracasse	130 – 410 mots	141 – 510 mots	Systématique

Tableau 1 : expérimentations chronométrées pour tester plusieurs stratégies d'annotation.

L'avant-dernière ligne du tableau fait apparaître la notion de chunk. En effet, pour valider nos décisions méthodologiques, nous avons également exploré une autre voie : celle de l'exploitation de pré-annotations (plutôt que de partir d'un texte nu).

L'aspect robotique du repérage des expressions référentielles (NB : pas de l'attribution d'un référent) peut être encouragé en utilisant un système de TAL en tant que pré-annotateur. Mais encore faut-il trouver un système de TAL qui soit adapté au français et dont le taux d'erreur n'entrave pas l'annotation : quand les erreurs sont nombreuses, l'annotateur a vite l'impression de passer son temps à les corriger plutôt qu'à exploiter directement les pré-annotations. Or il n'existe pas de système de détection automatique des expressions référentielles, et il nous faut nous rabattre sur des systèmes effectuant :

- la détection des entités nommées (ce qui est une tâche plus réduite[8]),
- la détection des anaphores (ce qui est également très réducteur),
- la détection des chaînes de référence – sauf que le seul système disponible pour la langue française, RefGen[9], présente des performances moyennes – ce qui justifie au passage les efforts fournis par le projet Democrat.

[8] Nouvel, D., Ehrmann, M., Rosset, S. (2015). *Les entités nommées pour le traitement automatique des langues.* Londres : Éditions ISTE.

En fin de compte, le système le plus proche du résultat souhaité est tout simplement un détecteur de chunks nominaux. Après un comparatif des performances des différents outils disponibles, notre choix a porté sur le chunker nominal de SEM[10].

Le principal avantage d'un chunker nominal est qu'il permet à l'annotateur de n'oublier aucune expression référentielle. Mais deux inconvénients importants viennent contrebalancer cet avantage :

1. tous les chunks nominaux d'un texte ne réfèrent pas, donc le chunker produit du bruit qui peut perturber l'annotateur (« il » impersonnel, mention non référentielle de partie du corps comme « avoir la grosse tête », etc.) ;
2. un chunker, par définition, repère des portions de texte non enchâssées – or des expressions référentielles peuvent s'enchâsser, comme les compléments du nom.

Des adaptations des résultats du chunker sont donc nécessaires et, là aussi, tout repose sur l'ergonomie de l'outil d'annotation utilisé. Ainsi, un outil qui permet de rectifier facilement les frontières d'un marquable peut avantager l'exploitation d'une pré-annotation en chunks.

Pendant un temps, quelques annotateurs du projet Democrat sont partis de textes nus pendant que quelques autres sont partis de textes chunkés. Au final, plus aucun annotateur ne se sert de textes chunkés. Autrement dit, la conclusion est la même que pour l'expérimentation consistant à filtrer les types de référents : à terme, on préfère une solution simple et sans aucun élément perturbateur ni questions à se poser tout au long de l'annotation (comme : « ce chunk est-il référentiel ou non ? »). **Les décisions méthodologiques de la section C2 sont donc validées**.

C.4 Mise en œuvre d'un schema d'annotation en fonction des outils

Deux questions se posent : comment gérer les expressions référentielles, et comment gérer les chaînes de référence. Pour la première, tous les outils d'annotation sont capables d'annoter des marquables et peuvent donc faire l'affaire. Pour la seconde, il existe deux possibilités principales, qui sont fonction des outils :

1. Relier une expression référentielle à l'expression précédente (ou à la première expression référant au même référent, comme dans le corpus ANCOR).
2. Construire un ensemble qui contient toutes les expressions référentielles portant sur le même référent.

Tous les outils permettant de gérer des relations autorisent la possibilité n° 1.

---

[9] Todiraşcu, A., Longo, L. (2011). RefGen, outil d'identification automatique des chaînes de référence en français. *18e Conférence sur le Traitement Automatique des Langues Naturelles, session des démonstrations industrielles*, Montpellier.

[10] Tellier, I., Duchier, D., Eshkol, I., Courmet, A., Martinet, M. (2012). Apprentissage automatique d'un chunker pour le français. In *Actes de la 19e Conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble.

Seuls les outils gérant des « schémas » (GLOZZ, ANALEC) permettent la possibilité n° 2.

Si nous avions choisi un outil gérant des relations mais pas des schémas, nous aurions dû suivre forcément la possibilité n° 1. Or, pour des raisons d'exploration du corpus et d'interrogation des données annotées, nous ne voulions pas exclure l'une de ces deux possibilités. C'est ainsi que nous avons choisi ANALEC (puis TXM). En effet, il ne faut pas oublier que l'annotateur humain annote des expressions référentielles, mais que c'est une macro qui construit les chaînes de référence (à partir des valeurs prises par le champ « REF »). Là où une macro peut être développée, on peut aussi en prévoir deux : une qui implémente la possibilité n° 1, l'autre qui implémente la possibilité n° 2 (et rend ainsi le corpus Democrat plus ou moins compatible avec le corpus ANCOR). **Nous avons donc décidé d'exploiter les possibilités des relations et des schémas**, et pas seulement des relations. Donc d'utiliser un outil fondé sur le modèle URS (unité, relation, schéma)[11].

C.5 Mise en place d'un manuel d'annotation

Une fois la structure des annotations décidée, il reste à écrire un manuel qui décrit à un annotateur quelconque (par exemple nouveau dans le projet Democrat) nos choix ainsi que la façon de les suivre quand on part d'un texte nu et que l'on souhaite obtenir un texte annoté.

Entre mai 2016 et février 2018, ce manuel a fait l'objet d'un document placé sur une plateforme de partage de fichiers (sharedocs d'Huma-Num). Le but était que chaque annotateur qui perçoit une imprécision ou une ambiguïté le signale – et, mieux : propose une correction. Plusieurs corrections se sont enchaînées. Le manuel a été stable pendant quelques mois, qui ont correspondu à l'annotation de textes littéraires. Puis des textes d'autres genres ont été annotés, notamment des textes de loi. Beaucoup d'exemples ne « collaient » pas avec ceux du manuel, ce qui a amené à revoir celui-ci. D'où une nouvelle série de corrections mi-2017. Les corrections n'ont pas cessé depuis, mais à un rythme cependant beaucoup moins soutenu. Pour information, la version qui constitue la section D de ce document correspond à la version 2.4.3 du manuel d'annotation. La voici.

D Manuel d'annotation

D.1 Introduction

D.1.1 Conventions et definitions

Une « expression référentielle », ou « mention », est une expression linguistique qui a un pouvoir de référence, c'est-à-dire qui renvoie à une entité du réel, de la fiction, etc. Chaque

[11] Widlöcher, A., Mathet, Y. (2012). The Glozz platform: A corpus annotation and mining tool. In: *Proceedings of the 2012 ACM Symposium on Document Engineering*, ACM, 171-180.

expression référentielle peut être matérialisée en tant que segment avec un début et une fin, symbolisés par des crochets dans les exemples :

[Le petit chat] boit [du lait].

Les expressions référentielles peuvent s'imbriquer les unes dans les autres :

[Le petit chat de [la voisine]] boit [du lait].

Dans la notation, on peut faire suivre les crochets par une lettre, appelée un *indice* (on commence traditionnellement par la lettre *i*, puis on continue l'alphabet : *j, k, l*, etc.). Lorsque deux expressions référentielles renvoient au même référent, elles ont le même indice :

[Le petit chat de [la voisine]$_j$]$_i$ boit [du lait]$_k$. **[Il]$_i$** s'appelle [Félix]$_l$.

Une expression qui est isolée, c'est-à-dire qui n'est reliée à aucune autre expression par un lien de coréférence, est appelée *« singleton »* (un ensemble à un seul élément).

Nous mettrons en gras les expressions qui illustrent le point discuté, comme dans l'exemple précédent. Nous noterons systématiquement *toutes* les expressions référentielles, même celles qui ne concerneront pas le point discuté, afin d'éviter que les silences ne soient interprétés de façon erronée : quand il n'y a pas de crochets, donc, ce n'est pas pour simplifier l'exemple et ne pas le surcharger, c'est réellement parce qu'il ne faut pas annoter l'expression.

D.1.2 Qu'est-ce que « l'annotation en coreference » ?

L'annotation comprend deux tâches principales :

- la délimitation des expressions référentielles (ce qui suppose de savoir ce qu'est une *expression référentielle*, où elle commence et où elle s'arrête). Il est important d'annoter *toutes* les expressions référentielles, même celles qui ne sont pas coréférentes,
- l'établissement des liens de coréférence (c'est-à-dire le fait de relier les expressions qui renvoient au même référent).

D.1.3 Les outils

Le projet Democrat permet à un annotateur d'utiliser au choix :

- TXM (http://textometrie.ens-lyon.fr) avec l'extension Analec (http://textometrie.ens-lyon.fr/html/doc/manual/manual68.html[12]),
- SACR (http://boberle.com/projects/sacr)[13],

[12] Voir également la version PDF du manuel : http://textometrie.ens-lyon.fr/files/documentation/Manuel%20de%20TXM%200.7%20FR.pdf.

[13] Oberlé, B. (2017) « Annotation of co-reference with SACR, a new 'drag-and-drop' tool », in: *ECLAVIT Workshop*, Université Paris-Est Marne-la-Vallée. https://halshs.archives-ouvertes.fr/halshs-01715467.

- Analec (http://www.lattice.cnrs.fr/-Analec-). C'est le premier outil à avoir été utilisé, au début du projet, et il est toujours possible de le faire. Néanmoins, son intégration dans TXM conduit à préférer ce dernier.

D.1.4 Choisir le nom du referent

On s'efforcera de choisir, pour les référents, une description singularisante (notamment un nom propre), permettant à l'annotateur, voire à un autre lecteur (correcteur), de comprendre de quel référent il s'agit. On n'inclura pas les possessifs (« ma mère »), les adjectifs et autres modifieurs qui ne sont pas utiles. Il est essentiel de conserver toujours le même nom pour le référent : la casse (différence entre majuscule et minuscule) et la ponctuation ont leur importance (« Pierre » est différent de « PIERRE » et de « pierre »).

Pour les singletons, il est possible d'utiliser le code spécial « SI » plutôt que de trouver un nom plus complexe. Ce code sera remplacé par un identifiant unique de manière automatique par la suite.

D.1.5 Attribution du nom du referent dans TXM

Dans TXM, après avoir créé une mention, il faut remplir le champ REF avec le nom qu'on a choisi pour le référent (cf. figure 1).

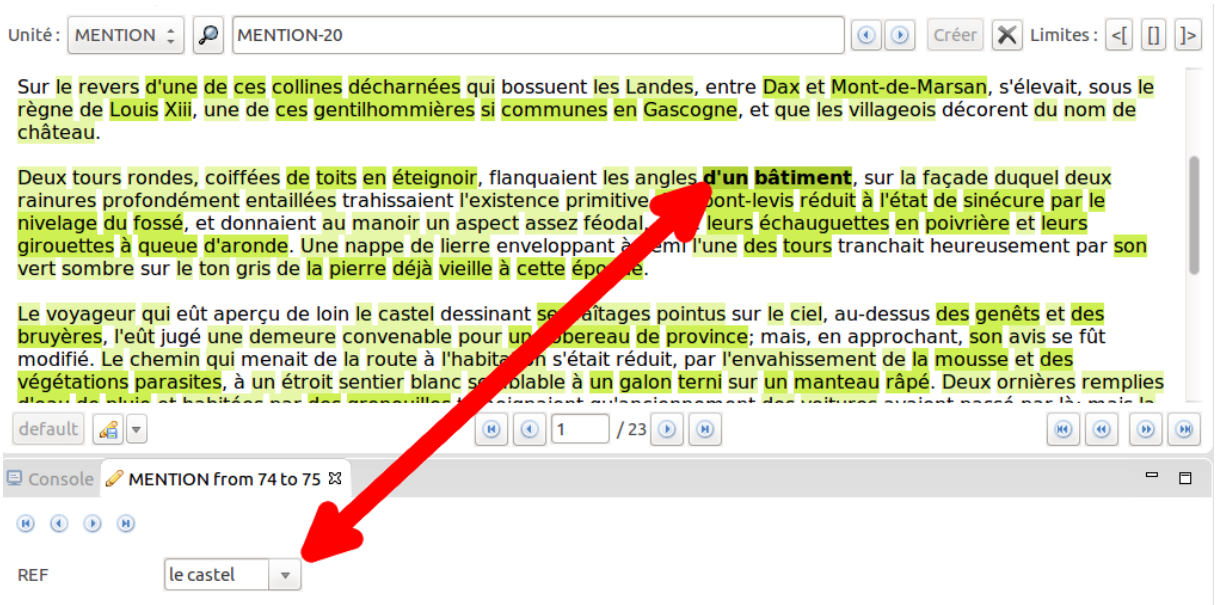

Figure 1 : annotation d'une mention, avec le champ (unique) « REF ».

Si une expression référentielle renvoyant à ce référent a déjà été annotée, il vaut mieux utiliser la liste déroulante pour retrouver le nom entré précédemment (cela évite les erreurs). La figure 2 montre ce procédé.

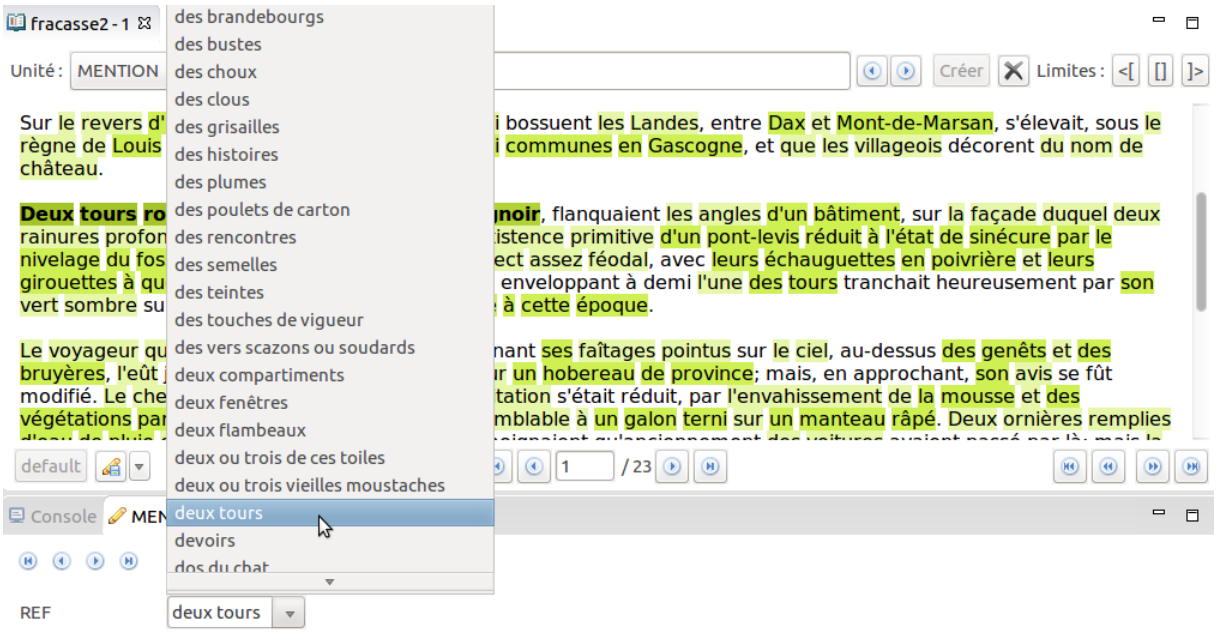

Figure 2 : annotation du champ REF en exploitant la liste de valeurs déjà saisie.

D.1.6 Attribution du nom du referent dans SACR

Il n'est pas indispensable de définir un nom de référent dans SACR (le fonctionnement de SACR est spécifique et vise directement la construction de chaînes), mais il faut le faire pour la campagne d'annotation Democrat. Lorsque l'on modifie le nom d'un référent, toutes les mentions liées à ce référent sont modifiées. Pour ce faire, il faut sélectionner une expression référentielle puis appuyer sur la touche « m » (comme « modifier »). Une boîte de dialogue avec un nom par défaut s'affiche, basé sur le contenu textuel de l'expression : on peut garder le nom proposé ou le modifier, puis appuyer sur OK. L'affichage se met à jour automatiquement, et toutes les expressions renvoyant au même référent ont désormais le même nom.

D.2 Annotation des expressions referentielles

D.2.1 Les noms

D.2.1.1 Quels noms annoter ?

D.2.1.1.1 Cas général

De manière générale, les noms, propres ou communs (ex. : *le petit chat, le chat, le même, ce chat*), sont susceptibles de référer : il faut donc les annoter :

[Paul]$_i$ mange **[la pomme]**$_j$.

Puisque les expressions référentielles peuvent s'imbriquer les unes dans les autres, on peut annoter des compléments du nom de compléments du nom, etc. :

[L'élection de [Macron]$_i$ à [la présidence de [la République]$_j$]$_k$]$_l$

D.2.1.1.2 Noms non référentiels

Certains noms, cependant, ne sont pas référentiels :

- les noms inclus dans des expressions figées comme *jouer **des coudes*** ou *au **coup** par **coup***,
- les noms communs épithètes (c'est-à-dire qui sont directement apposés à un autre nom) : *[un roman **fleuve**]* ; *[un sujet un peu **bateau**]*
- les noms propres épithètes qui ont une valeur identifiante : *[le prix **Jean Moulin**] ; [le prix **Nobel**] ; [l'affaire **Dreyfuss**] ; [le style **Louis XVI**] ; [l'effet **Balladur**]*
- les noms sans déterminant introduits par une préposition s'ils forment un nom composé, comme dans *[armoire à **glace**]* (surtout si l'expression désigne par métaphore un homme fort)

Attention : Les noms, notamment géographiques, qui entrent dans la composition de *certaines* entités nommées (c'est-à-dire des expressions désignant des personnes, des organisations, des lieux, des dates, des quantités, etc.), sont annotés en tant qu'expression référentielle :

> [L'espace **Schengen**]$_i$ comprend [les territoires [des États]$_j$ [qui]$_j$ ont mis en œuvre [l'accord de **[Schengen]$_k$**]$_l$]$_m$.

D.2.1.1.3 Le cas particulier des noms sans déterminant

Bien que les noms communs sans déterminant n'aient souvent pas vocation à référer, nous les annotons, car ils peuvent être repris :

> [La plupart du temps]$_i$, [je]$_j$ dors en **[cours]$_k$**.
>
> [[Les chefs de **[départements]$_i$**]$_j$…
>
> [Elle]$_i$ était sans **[enfant]$_j$**. Pourtant [elle]$_i$ **[en]$_j$** aurait voulu.
>
> [J']$_i$écris généralement en **[bleu]$_j$**. **[Cette couleur]$_j$** est jolie.

Nous ne les annotons pas, cependant, s'ils sont dans une expression figée (voir section D.2.1.1.4), comme *à condition de,* sauf s'ils sont repris. Comparer :

> [Il]$_i$ aura [un bonbon]$_j$ à **condition** qu'[il]$_i$ travaille bien.
>
> [Il]$_i$ aura [un bonbon]$_j$ à **[condition]$_k$** qu'[il]$_i$ travaille bien. C'est à **[cette condition]$_k$** seulement qu'[il]$_i$ sera récompensé.

D.2.1.1.4 Les locutions prépositionnelles et conjonctives

Certaines locutions prépositionnelles et conjonctives comme *en cours de, en matière de, à condition de, aux fins de, à l'exception de, à l'exclusion de, en cas de,* etc. comportent des noms qui ne sont pas référentiels. On ne les annote pas, sauf s'ils sont repris :

> [J']$_i$aime [tous les fruits, à l'exception [des tomates]$_j$]$_k$.

[J']$_i$aime [tous les fruits, à [l'exception]$_l$ [des tomates]$_j$]$_k$. [Cette exception]$_l$ n'est pas étonnante : [les tomates]$_j$ ne sont généralement pas considérées comme [des fruits]$_m$.

Par contre, on annotera les noms dans les expressions qui ne sont pas vraiment figées, comme *dans le désir de*, *dans le but de*, etc. :

Dans **[le désir]$_i$ de** garder [l'inflation]$_j$ sous [contrôle]$_k$...

D.2.1.1.5 Les termes d'adresse

Les termes d'adresse (vocatifs) sont annotés :

[Cher ami]$_i$, qu'[en]$_j$ pensez [vous]$_i$?

D.2.1.1.6 Les nombres et les dates

Les nombres ne sont annotés que lorsqu'ils renvoient à une date. Celles-ci sont annotées comme une seule expression référentielle :

[Le 1er juin 2017]$_i$

En **[1968]$_i$**, [je]$_j$ n'étais pas né.

En **[mars]$_i$**, [le printemps]$_j$ revient.

[Charlemagne]$_i$ a été couronné [empereur]$_i$ **[le 25 décembre 800 après [Jésus-Christ]$_j$]$_k$**

D.2.1.1.7 Les attributs

Tous les attributs nominaux (pas les adjectifs) sont annotés comme des expressions référentielles :

[Marie]$_i$ est **[vendeuse dans [une boulangerie]$_j$]$_i$**.

[Londres]$_i$ est **[la capitale de [l'Angleterre]$_j$]$_i$**.

[Louis]$_i$ était en réalité **[ce fameux colonel]$_j$** [dont]$_j$ [il]$_k$ avait tant entendu parler.

Les attributs peuvent aussi être introduits par d'autres verbes que *être* :

[Il]$_i$ est considéré comme **[empereur]$_i$**.

[Il]$_i$ a été couronné **[empereur]$_i$**.

[L'aventure]$_i$ [lui]$_j$ paraissait **[un heureux présage]$_i$**.

Attention : il y a des différences dans la gestion de la coréférence (voir la section D.3.7).

D.2.1.1.8 En mention, en usage

Certains mots peuvent être utilisés *en mention* (par opposition à « en usage »), lorsqu'il est fait référence au nom en tant que tel. Dans ce cas, le référent est le mot lui-même, comme dans :

[*Pomme*]$_i$ s'écrit avec [deux *m*]$_j$. [C']$_i$est [un mot complexe]$_i$.

Il n'y a pas coréférence entre l'objet et le terme, comme le montrent les deux exemples suivants :

[Paul]$_i$ est [jardinier]$_i$… [Il]$_i$ a un beau [prénom]$_j$.

On aura donc :

[Paul]$_i$, nommé **[Pierre]$_j$** chez [les Grecs]$_k$...

[Il]$_i$ s'appelle **[Charlemagne]$_j$**.

[On]$_i$ [l']$_j$ a appelé **[Charlemagne]$_k$**.

De même :

[Protocole portant [modification de [la Convention pour [l'unification de [certaines règles relatives [au transport aérien international]$_i$]$_j$]$_k$ signée à [Varsovie]$_l$ [le 12 octobre 1929]$_m$]$_n$]$_o$, fait à [La Haye]$_p$ [le 28 septembre 1955]$_q$]$_r$ (appelé ci-après **[le Protocole de [La Haye]$_p$]$_s$**)

C'est également le cas de reformulation de noms :

[Sardan-Pul]$_i$, c'est à dire, [Sardan fils de [Phul]$_j$]$_k$

et des expressions étrangères :

[Le renard ordinaire]$_i$, [*vulpes vulgaris*]$_j$, Boit. ; [*canis vulpes*]$_k$, Lin. ; [le *fuchs* [des Allemands]$_l$]$_m$; [le *fox* [des anglais]$_n$]$_o$; [le *raf* [des Suédois]$_p$]$_q$; [le *zorra* [des Espagnols]$_r$]$_s$...

D.2.1.1.9 Les titres et les listes numérotées

Les titres peuvent désigner soit la structure du document (« chapitre 3 »), soit le contenu (« Les entités nommées »), soit les deux (« Paragraphe 3 relatif au transport des glaçons en cas de canicule »). On annote les éléments du titre :

[Chapitre 3]$_i$

[Les entités nommées]$_i$

[Paragraphe 3 relatif [au transport [des glaçons]$_i$]$_j$ en cas de [canicule]$_k$]$_l$

[Paragraphe 3]$_i$: [Le transport [des glaçons]$_j$]$_k$ en cas de [canicule]$_l$

On annote aussi les numéros des listes numérotées, s'ils sont matérialisés par des lettres ou des chiffres :

[a)]$_i$ [transport par [camion réfrigéré]$_k$]$_j$

[b)]$_l$ [transport en [glacière]$_n$]$_m$

... comme évoqué dans [le point (a)]$_i$...

D.2.1.1.10 Les déterminants complexes

Un déterminant complexe peut être formé par une coordination entre deux déterminants simples. Dans ce cas, on annote comme s'il s'agissait d'un seul déterminant :

> [Une ou plusieurs escales]$_i$...

Certains noms qui évoquent des contenants (*bouteille, camion, brouette*) peuvent soit être utilisés en tant que nom référentiel, soit former des déterminants complexes (on les appelle alors des *substantifs quantificateurs*). Dans un souci de simplicité, nous les annotons dans tous les cas :

> [La bouteille d'[eau]$_j$]$_i$ est tombée par terre. [Elle]$_i$ s'est cassée en [mille morceaux]$_k$.
>
> [J']$_i$ai bu [une bouteille d'[eau]$_j$]$_k$, mais [l'eau]$_j$ n'était pas bonne.

D.2.1.2 Comment délimiter une expression référentielle nominale ?

D.2.1.2.1 Les modifieurs inclus dans l'expression

Democrat n'annote pas des syntagmes entiers, mais des « bouts » de syntagmes, ce qui rend la délimitation des expressions référentielles délicate. Les syntagmes nominaux simples, composés d'un déterminant et d'un nom, avec éventuellement un adjectif antéposé, ne posent pas de problème :

> [Le petit chat]$_i$

Ce sont les modifieurs post-posés auxquels il faut faire attention, car seuls quelques-uns sont inclus dans la mention :

- les adjectifs : *[Le chat **noir**]$_i$*
- les compléments du nom, quelle que soit la préposition :
  - [L'élection **de [Macron]$_i$**]$_j$
  - [J']$_i$ai [les compétences **pour vendre [des glaces]$_j$**]$_k$ sur [la plage **à côté de [Biarritz]$_l$**]$_m$.
  - [Les rapports **entre [[l'expéditeur]$_i$ et [le destinataire]$_j$]$_k$**]$_l$
  - [L'intérêt [des écrivains]$_i$]$_j$
  - [L'élection [du président de [la République]$_i$]$_j$ **par [le peuple]$_k$**]$_l$

    Avec ce dernier exemple, on voit qu'on inclut tous les arguments de la structure argumentale d'un prédicat nominal.
- les noms épithètes, aussi appelés appositions liées : *[Le prix Jean Moulin] ; [[Leur]$_i$ grand-père Numitor]$_j$* (voir la section D.2.1.2.3 ci-dessous sur les appositions)

- les participes, passés ou présents, avec toute leur structure argumentale : *[Le président **élu**]$_i$; [Le président **élu par le peuple français avec 66,1 % des voix**]$_i$*

  sauf s'ils sont apposés (voir la section D.2.1.2.3) : *[Le président]$_i$, élu,... ; [Le président]$_i$, élu par le peuple français avec 66,1 % des voix,...*

Quand il y a plusieurs modifieurs, on les inclut tous dans la mention, quel que soit leur catégorie grammaticale d'appartenance :

[Le chat noir [au miaulement étrange]$_i$]$_j$

[Le chat noir et roux]$_i$

[des petits robots alertes, faciles à [l'épouvante]$_j$ et prompts à [l'enthousiasme]$_k$, prêts à [l'attaque]$_l$ comme à [la fuite]$_m$]$_i$

D.2.1.2.2 Les relatives

Les relatives ne sont pas comprises dans les limites de l'expression référentielle. Mais le pronom relatif est annoté :

[Le petit chat]$_i$ **[qui]$_i$ boit [du lait]$_j$**...

Pour les pronoms relatifs sans antécédents, voir la section D.2.2.3.

D.2.1.2.3 Les appositions

Les appositions, qu'elles soient adjectives, relatives ou nominales, ne figurent pas dans une mention. Il est parfois difficile de savoir si tel terme est en apposition ou non, surtout pour les états anciens de la langue. Pour simplifier, on considérera qu'il y a apposition dès lors qu'il y a virgule :

[Le chat]$_i$, **petit**, boit [du lait]$_j$.

[Le chat]$_i$, [**qui]$_i$ est petit**, boit [du lait]$_j$.

[Macron]$_i$, **(le) président de** [**la République**]$_j$, a prononcé [un discours]$_k$.

[Achaz roy de [Juda]$_i$]$_j$

[Le roy Ochosias fils de [Joram roy de [[Juda]$_i$ et d'[Athalie]$_j$]$_k$]$_l$]$_m$

Dans l'exemple suivant, on inclut tout dans la même mention, considérant que la coordination exprimée par *et* l'emporte sur la virgule :

[Jézabeth soeur d'[Ochosias]$_i$, et femme de [Joïada souverain pontife]$_j$]$_k$

mais comparer avec :

[Jézabeth soeur d'[Ochosias]$_j$]$_i$, femme de [Joïada souverain pontife]$_k$

[Joïada]$_l$, souverain pontife

Cependant, quand une mention se trouve discontinue, sous l'effet de segment(s) apposé(s), comme dans l'exemple suivant, l'apposition sera annotée dans la mention pour éviter les mentions discontinues ou « trouées » :

[L'exploitation, **effective ou visée**, [des stocks]$_i$]$_j$

On fera aussi attention au fait que les structures énumératives d'adjectifs, de noms, etc. dont les unités sont séparées par des virgules, ne sont pas à considérer comme des cas d'apposition :

[Une voix **forte, caverneuse et rauque]$_i$…**

[Une voix forte, **mais caverneuse et rauque**]$_i$...

D.2.1.2.4 « Tel que »

Le traitement de « N_1 tel(le) que Pro/N_2 » est particulier : N_1 forme une expression référentielle à part, tout comme Pro/N_2, et les deux expressions sont coréférentes :

[Je]$_i$ ne suis pas surpris qu'**[un homme]$_j$** tel que **[sir George Burnwell]$_j$** ait exercé [une si profonde influence]$_k$ sur [lui]$_l$.

[Paul]$_i$ va s'engager dans l'armée. [Il]$_i$ [en]$_j$ a toujours rêvé. **[Un homme]$_i$** tel que **[lui]$_i$** fera [un bon soldat]$_i$.

D.2.1.2.5 Les prépositions

Les prépositions ne figurent pas dans les expressions référentielles, sauf dans le cas des amalgames :

[Clovis]$_i$ était [le roi [**des** Français]$_j$]$_k$.

[Le bord **de** [la table]$_i$]$_j$.

[Le bord [**du** bureau]$_i$]$_j$.

D.2.1.2.6 Les parenthèses

Le contenu des parenthèses est annoté, mais les parenthèses elles-mêmes ne sont pas incluses dans la mention :

[Le petit chat]$_i$ **([qui]$_i$ est noir)** boit [du lait]$_j$.

[L'UMP]$_i$ ([Union pour la Majorité Présidentielle]$_j$) a changé de nom.

D.2.2 LES PRONOMS

D.2.2.1 Quels pronoms annoter ?

Les pronoms sont en général annotés. C'est notamment le cas :

- des pronoms personnels (*moi, toi, lui, elle, nous, vous, eux, elles, le tien, le mien, moi-même,* etc., ce qui inclut aussi les clitiques sujets : *je, tu, il, elle, on, nous, vous, ils, elles* ; et les clitiques objets : *me, te, le, la, les, lui, leur, y, en*) : ***[Il]$_i$*** *est parti.*
- des pronoms démonstratifs : ***[Celui-ci]$_i$*** *est bleu.*
- des pronoms relatifs (qu'ils introduisent des relatives déterminatives ou explicatives), par exemple *qui, que, quoi, dont, où, lequel, quiconque* : *[Le village]$_i$* ***[qui]$_i$*** *est dans [la montagne]$_j$.*
- des pronoms dits adverbiaux (par exemple *où*) : *[Le village]$_i$* ***[où]$_i$*** *[j']$_k$ai grandi.*
- les autres pronoms (par exemples les indéfinis : *quelques-uns, plusieurs…* ou les interrogatifs : *qui, que, quoi, lequel*...), sauf ceux cités ci-après.

Certains pronoms ne doivent cependant pas être annotés :

- les pronoms réfléchis clitiques (c'est-à-dire quand ils sont *clitiques* et *renvoient au sujet de la proposition*) : *[Je]$_i$* ***me*** *lave.* (et non : *[Je]$_i$* ***[me]$_i$*** *lave.*)

  Par contre, les pronoms clitiques objets qui ne sont pas réfléchis doivent être annotés : *[Il]$_i$* ***[le]$_j$*** *lave.* (à ne pas confondre avec : *Il* ***se*** *lave.*)

  De même, les réfléchis non clitiques doivent être annotés : *[Il]$_i$ se lave* ***[lui-même]$_i$****. [Il]$_i$ se donne [un cadeau]$_j$ à* ***[lui-même]$_i$****.*
- les pronoms qui figurent dans des expressions figées (notamment verbales) : *[Il]$_i$ se* ***la*** *joue un peu trop ; [Il]$_i$ se* ***la*** *coule douce ; [Il]$_i$* ***en*** *fait des tonnes.*
- les pronoms impersonnels : ***Il*** *pleut.* ***Il*** *manque de [la glace]$_i$.*
- les pronoms négatifs, c'est-à-dire qui désignent qu'il n'y a aucun élément : *[[Pierre]$_i$ et [Marie]$_j$]$_k$ se sont mariés.* ***Aucun [des deux]$_k$*** *n'a pensé à commander [un gâteau]$_l$.* ***Aucun [des deux]$_k$*** *non plus n'a pensé à faire venir [un traiteur]$_m$.*

D.2.2.2 Le pronom « l'on »

On annote « l'on » d'un seul bloc :

> [Je]$_i$ pense que [l'on]$_j$ peut dire...

D.2.2.3 Les pronoms sans antécédent

Généralement, la délimitation de l'expression pronominale ne pose pas de problème. Cependant, lorsqu'un pronom démonstratif n'a pas d'antécédent et que sa référence se construit par une relative qui le suit, on annote l'ensemble comme une seule mention :

> [Il]$_i$ a **[tout ce que [l'on]$_j$ peut attendre d'[un conseiller]$_k$]$_l$**.
>
> [Deux hommes]$_i$ étaient devant [lui]$_j$. **[Celui qui avait [la sacoche]$_l$]$_k$** s'appelait Malcolm.

Dans ce cas, le pronom relatif *qui/que* ne constitue pas une mention à part, et toute la relative est annotée.

Même chose pour le « pronom » *un* suivi d'un relatif :

Il y [en]$_i$ a **[un qui dort]$_j$**.

Il en va de même pour les relatifs sans antécédent :

[Qui dort]$_i$ dîne.

D.2.2.4 « C'est »

Dans le cas de « c'est ADJ », le *c'* n'est pas référentiel : on ne l'annote pas.

Dans le cas de « c'est (DET) N », on annote *c'* et le N ; les deux coréfèrent :

[C']$_i$était [des hordes débandées]$_i$.

[Un homme]$_i$ est venu. [C']$_i$était [Paul]$_i$.

« C'est » apparaît également dans les phrases clivées et pseudo-clivées. Dans le cas du clivage, on n'annote ni *c'* ni *qui* :

C'est [la linguistique]$_i$ **qui** [m']$_j$intéresse.

C'est [Charles]$_i$ **qui** mange.

Dans le cas de la pseudo-clivée, par contre, on annote de la façon suivante :

[Ce qui [m']$_i$intéresse]$_j$, [c']$_j$est [la linguistique]$_j$.

[Celui qui mange]$_i$, [c']$_i$est [Charles]$_i$

Dans « c'est que », le *c'* n'est pas référentiel :

C'est que [je]$_i$ suis [le roi]$_i$.

D.2.3 LES VERBES

D.2.3.1 Les sujets zéros

Les verbes ne sont annotés qu'en l'absence de sujet exprimé. Ils portent alors l'indication qu'ils ont un sujet zéro.

D.2.3.1.1 Coordination

Il y a sujet zéro dans le cas d'une coordination :

[Pierre]$_i$ boit et **[fume]$_i$**.

Les adverbes (notamment de négation) ne sont pas inclus dans les mentions :

[Pierre]$_i$ boit mais ne **[fume]$_i$** pas.

D.2.3.1.2 Impératif

Il y a aussi sujet zéro dans le cas de l'impératif :

> **[Ferme]$_i$** [la porte]$_j$.

D.2.3.2 Les verbes support d'une anaphore

Certains prédicats, bien qu'à la source d'anaphores, ne sont pas annotés :

> [Les vêtements récoltés]$_i$ vont être **vendus** à [la société Euro-collecte]$_j$. [Le produit]$_k$ de **[cette vente]$_l$** sera reversé à…
>
> Il **neige**. **[Elle]$_i$** tient.

D.2.4 LES DETERMINANTS

Les déterminants ne sont pas référentiels : on ne les annote pas. Seul le déterminant possessif (*mon, ton son, ma, ta, sa, mes, tes, ses, notre, votre, leur, nos, vos, leurs*) fait exception, car il encode le possesseur de sorte que *Marie... son chien* se comprend comme *Marie... le chien d'elle (= Marie).* On établit donc un lien de coréférence entre le déterminant possessif et le possesseur :

> [Jean]$_i$ a apprécié [**[son]$_i$** café]$_j$.
>
> [**[Mon]$_i$** départ]$_j$

Attention : les possessifs des expressions figées, comme ***ma** foi, **mon** dieu, j'en perds **mon** latin*, etc. ne doivent pas être annotés.

D.2.5 AUTRES CAS

D.2.5.1 Les groupes

Lorsque deux ou plusieurs noms sont coordonnés, il faut annoter chacun de ces noms, mais aussi le groupe entier, car il peut être repris :

> **[[Pierre]$_i$ et [Paul]$_j$]$_k$** s'aiment. **[Ils]$_k$** vont se marier.
>
> [J']$_i$ai vu [au zoo]$_j$ **[[un lion]$_k$, [un zèbre]$_l$ et [un singe]$_m$]$_n$**. **[Ils]$_n$** avaient l'air tristes.

Différents coordonnants peuvent être utilisés : *avec, ainsi que, aidé de*, etc. Par contre, *ou* n'entraîne pas la création d'un groupe si l'alternative est exclusive, c'est-à-dire si le verbe reste au singulier :

> [[Pierre]$_i$ **et** [Marie]$_j$]$_k$ **iront** faire [les courses]$_l$.
>
> [Pierre]$_i$ **ou** [Marie]$_j$ **ira** faire [les courses]$_k$.

Parfois, une expansion s'applique à chaque élément du groupe, mais n'est pas répétée (elle est mise *en facteur commun*). Dans ce cas, elle est rattachée au dernier nom :

> [[L'achat]$_i$ et [la démolition d'**[une maison]$_j$**]$_k$]$_l$

Ici, même si « d'une maison » est une expansion qui s'applique à « l'achat » et à « la démolition », elle n'est rattachée qu'au deuxième nom.

Certaines prépositions indiquent l'inclusion ou l'exclusion d'un sous-ensemble :

> [Les fruits, **y compris [les tomates]$_i$**]$_j$, contiennent [des graines]$_k$.

> [Les fruits, **à l'exception [des tomates]$_i$**]$_j$, se mangent en [dessert]$_k$.

Quand les antécédents sont *dispersés* de part et d'autre d'un prédicat (cas dits de *split antecedent*), ils font l'objet de mentions séparées :

> **[Pierre]$_i$** retrouva **[[sa]$_i$ femme]$_j$** au restaurant. [Ils]$_k$ dînèrent jusqu'à tard dans [la nuit]$_l$.

D.2.5.2 Le cas des anaphores résomptives

Un nom ou un pronom (souvent démonstratif) peut reprendre une phrase, un paragraphe ou même toute une partie du texte : c'est une *anaphore résomptive*. On n'annote pas le segment de texte :

> [Il]$_i$ a dit qu'[il]$_i$ avait brisé [le vase]$_j$ en trébuchant. **[Cette explication]$_k$** est cependant douteuse : **[elle]$_k$** a été contredite par [[son]$_i$ frère]$_l$.

> [Il]$_i$ a dit qu'[il]$_i$ avait brisé [le vase]$_j$ en trébuchant. **[Cela]$_k$** est faux.

D.3 QUESTIONS DE COREFERENCE

D.3.1 ANAPHORE ASSOCIATIVE ET AUTRES LIENS SEMANTIQUES

La partie n'est pas équivalente au tout, il n'y a donc pas de coréférence :

> [Le veau]$_i$… **[[sa]$_i$ tête]$_j$**.

> [[Son]$_i$ calendrier]$_j$... **[Les quatre premières colonnes]$_k$**…

De même, on n'annote pas par un lien de coréférence les relations sémantiques illustrées par l'exemple suivant :

> [Je]$_i$ fus mise en **[nourrice]$_j$**, dans [la banlieue de [Paris]$_k$]$_l$. C'est là que germa et grandit [cette passion pour **[celle]$_m$** à [qui]$_m$ [on]$_n$ [m']$_i$avait confiée]$_o$.

D.3.2 REFERENCE GENERIQUE, SPECIFIQUE, PARTICULIERE

Tous les termes renvoyant à une même référence générique sont considérés comme coréférents, même s'il y a passage du singulier au pluriel ou changement de déterminant :

[Les chats]$_i$ miaulent. [Un chat]$_i$, [ça]$_i$ miaule. [Le chat]$_i$, quand [il]$_i$ a faim, miaule. Etc.

Mais un référent générique n'est pas coréférent avec un référent spécifique ou particulier :

[Il]$_i$ a pris **[un café]$_j$**. **[Un café]$_k$,** [ça]$_k$ réchauffe toujours.

[Elle]$_i$ ne pouvait donc s'embarrasser d'**[un enfant]$_j$**, et **[je]$_k$** fus mise en [nourrice]$_l$.

D.3.3 Discours direct

On ne considère pas qu'il s'effectue une rupture de la chaîne entre le passage du discours direct au discours indirect et inversement. Ce qui prime est l'information fournie au lecteur :

[Jeanne]$_i$ répondit : « Entre, **[papa]$_j$**. » Et **[[son]$_i$ père]$_j$** parut.

D.3.4 Termes equivalents

Il y a coréférence lorsque plusieurs expressions linguistiques renvoient au même référent. Cela ne pose généralement pas de problème pour les référents concrets :

[Le chat]$_i$… [Cet animal]$_i$… [Félix]$_i$…

Mais pour les entités abstraites, les choses sont plus compliquées : l'assassinat et la mort de César sont-ils le même référent ? L'armée française est-elle la même entité que l'ensemble des militaires français ? Ou bien que l'ensemble des soldats français ? L'annotation qui va dans le sens de la coréférence doit être privilégiée.

D.3.5 Referents flous

Parfois, il est difficile d'établir quel est le référent d'une expression référentielle.

C'est le cas par exemple de *on*. Il peut désigner un groupe précis incluant le locuteur (*Pierre et moi sommes allés au cinéma. Ensuite, **on** a fait la tournée des bars*), un référent générique (***on** dit que Trump est président des États-Unis*), un groupe défini mais non identifié (***on** m'a dit que Trump est président des États-Unis*), etc. On s'efforcera de faire la distinction dans chaque cas.

Avec *l'un, l'autre, chacun*, etc., il est parfois difficile de savoir à qui ou quoi le pronom réfère (« J'ai envie de me faire saltimbanque », disait **l'un**). Si l'ambiguïté ne peut pas être levée, on créera un référent à part « A ou B ».

D.3.6 REFERENTS EVOLUTIFS

Un référent peut se transformer au cours du texte. C'est le cas du *poulet bien vivant* et qu'on découpe, qu'on passe au four et qu'on sert bien chaud (à ce moment-là, il n'est normalement plus « bien vivant »). On considère dans ce cas que toutes les mentions sont coréférentes. Dans l'exemple suivant, le loup se transforme en « mère-grand » (le premier *la* renvoie au Petit Chaperon rouge ; nous n'annotons que les mentions du loup) :

> [Le Loup]$_i$, la voyant entrer, lui dit en se cachant dans le lit, sous la couverture : Mets la galette et le petit pot de beurre sur la huche, et viens te coucher avec [moi]$_i$. Le petit Chaperon rouge se déshabille, et va se mettre dans le lit, où elle fut bien étonnée de voir comment [sa mère-grand]$_i$ était faite en [son]$_i$ déshabillé. — Elle [lui]$_i$ dit : [Ma mère-grand]$_i$, que [vous]$_i$ avez de grands bras !

Ou encore :

> [L'ogre]$_i$ se changea en [une souris]$_i$, [qui]$_i$ se mit à courir sur [le plancher]$_j$.

D.3.7 LA COREFERENCE DES ATTRIBUTS

Les attributs posent des problèmes particuliers, parce qu'il y en a de différents types. On rappelle que *tous* les attributs nominaux sont annotés comme des expressions référentielles (voir la section D.2.1.1.7). Mais tous ne coréfèrent pas forcément au sujet de la phrase (ou à l'objet, pour les attributs de l'objet). De façon générale, les attributs du sujet (objet) coréfèrent avec le sujet (objet) :

> [Marie]$_i$ est **[vendeuse dans [une boulangerie]$_j$]$_i$**. [Elle]$_i$ est aussi **[la gérante d'[une pizzeria]$_k$]$_i$**.
>
> [Londres]$_i$ est **[la capitale de [l'Angleterre]$_j$]$_i$**.

Cependant, lorsque deux chaînes ont été construites indépendamment l'une de l'autre et se trouvent reliées dans une construction attributive (c'est-à-dire lorsqu'un même référent a deux noms différents, et qu'une construction attributive vient ensuite asserter l'égalité entre les deux noms), on ne fait pas le lien de coréférence :

> [J']$_i$aime regarder [l'étoile du [matin]$_j$]$_k$. [[Mon]$_i$ frère]$_l$ préfère regarder [l'étoile du [soir]$_m$]$_n$. [Nous]$_o$ savons pourtant que **[l'étoile du [matin]$_j$]$_k$ est [l'étoile du [soir]$_m$]$_n$**.

D.4 MODIFICATION DE LA STRUCTURE D'ANNOTATION AVEC TXM

Il ne faut ni modifier ni supprimer le champ REF dans la structure d'annotation. Cependant, vous pouvez y ajouter des propriétés et des valeurs en fonction de vos besoins d'annotation.

D.4.1 Ouverture de la structure

Un clic-droit sur le corpus permet d'accéder au menu pour ouvrir la structure d'annotation (cf. figure 3).

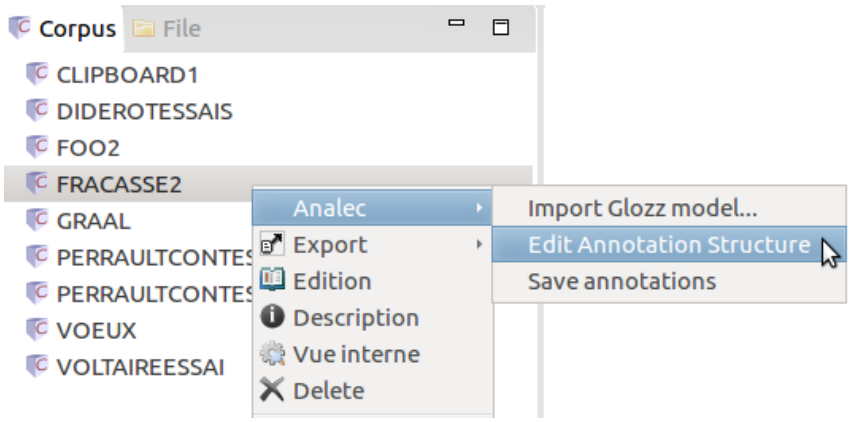

Figure 3 : accéder à la structure d'annotation dans TXM.

D.4.2 Ajout d'une propriete

Il suffit de faire un clic-droit sur le type pour lui ajouter une propriété (cf. figure 4).

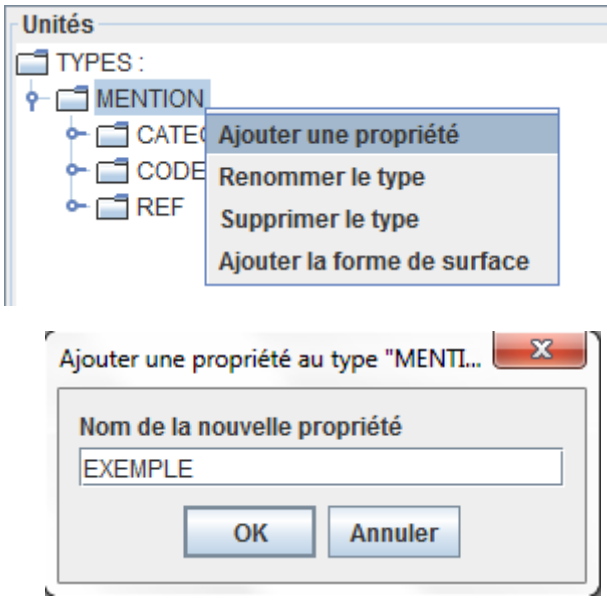

Figure 4 : ajout d'une propriété à un type d'unité dans la structure d'annotation.

D.4.3 Ajout d'une valeur

Faites un clic-droit sur la propriété pour lui ajouter une valeur (cf. figures 5a et 5b).

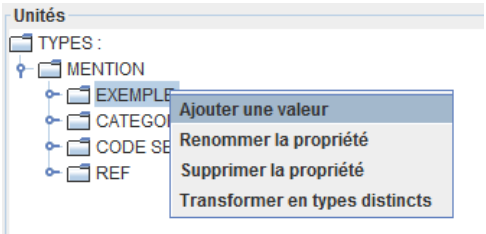

Figure 5a : annotation d'une valeur à une propriété dans la structure d'annotation (1/2).

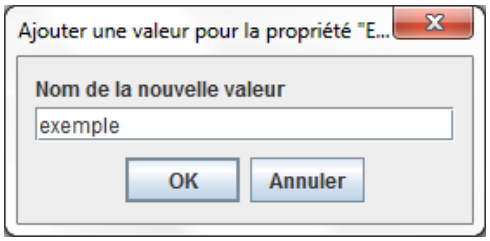

Figure 5b : annotation d'une valeur à une propriété dans la structure d'annotation (2/2).

La structure modifiée apparaît dans la figure 6.

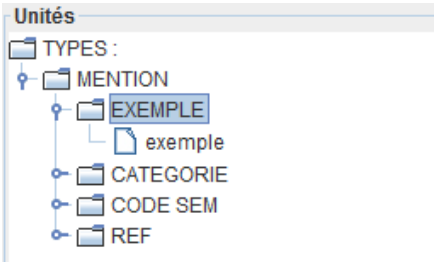

Figure 6 : structure d'annotation obtenue après ces deux ajouts.

D.4.4 Renommer ou supprimer les ajouts a la structure

Faites un clic-droit sur l'élément visé pour le renommer ou le supprimer (cf. figure 7).

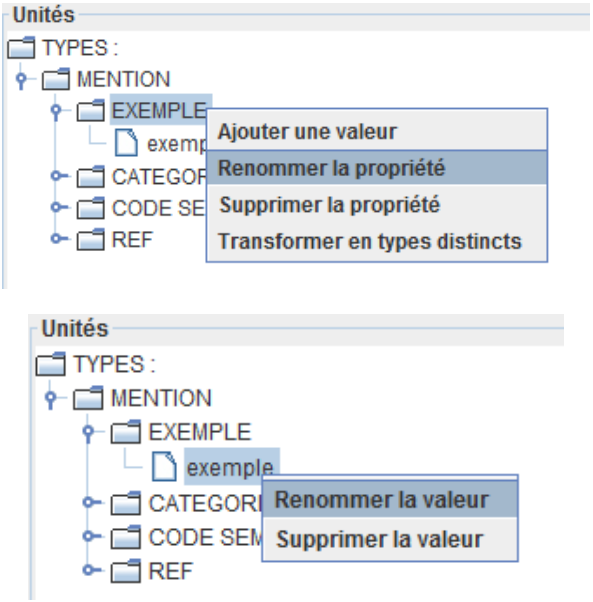

Figure 7 : accès au menu contextuel pour renommer une propriété ou une valeur d'une propriété.

D.4.5 Correction de la valeur du champ REF

Si vous souhaitez corriger la valeur du champ REF après une erreur de saisie (deux mentions ont un référent différent au lieu d'un référent commun), il faut renommer les valeurs en question.

Par exemple, « mère du narrateur » et « la mère du narrateur » renvoient au même référent et l'une des deux mentions doit être modifiée (cf. figures 8, 9 et 10).

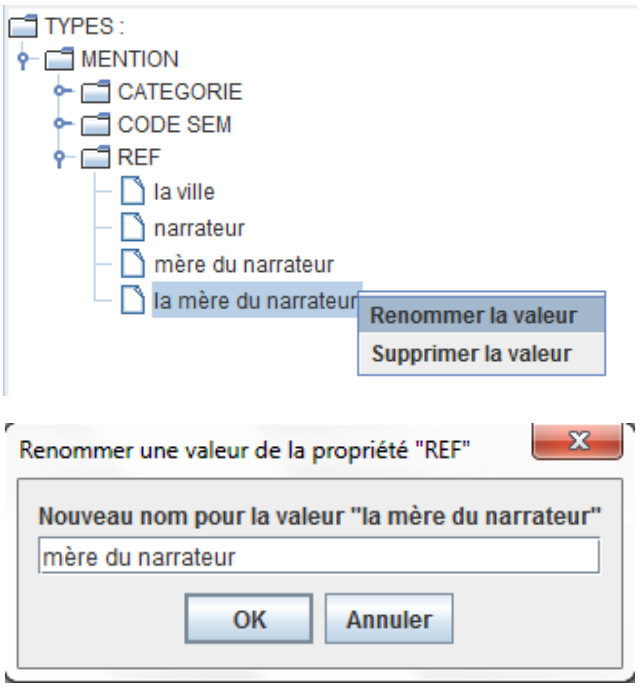

Figure 8 : renommage d'une valeur de propriété avec une valeur déjà existante.

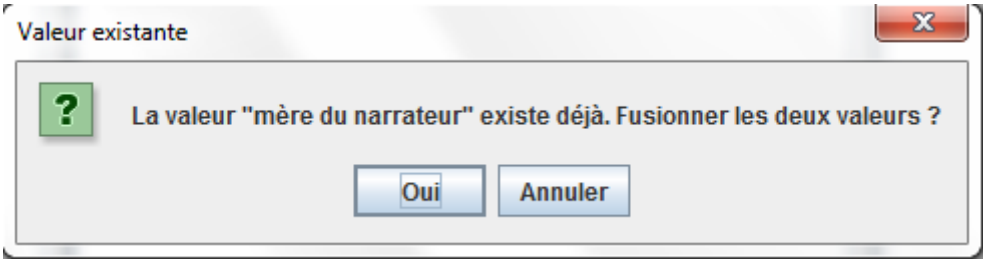

Figure 9 : réaction de TXM : proposition de fusionner les annotations ayant la valeur existante et celles ayant la valeur renommée.

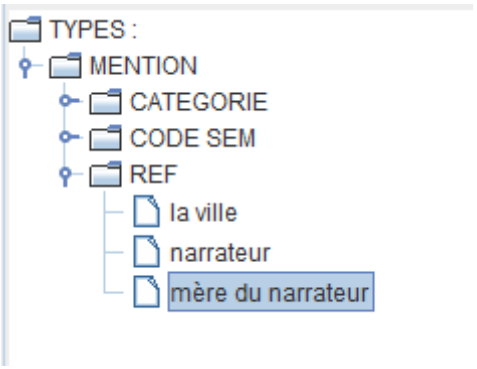

Figure 10 : résultat.

D.5 CONSTRUCTION AUTOMATIQUE ET ANNOTATION DES CHAINES DE REFERENCE

D.5.1 PRINCIPE

Une fois l'annotation des mentions terminées, il s'agit de s'occuper des chaînes de référence. Pour cela, TXM proposera en temps voulu une macro permettant la construction automatique de chaînes, chaque chaîne faisant l'objet d'un schéma (au sens du modèle URS – unités,

relations, schémas – utilisé pour les fonctionnalités d'annotation des outils Glozz, Analec et désormais TXM). A l'heure actuelle, les efforts ont porté sur l'annotation des mentions, et l'annotation des chaînes n'a pas encore été réalisée sur la totalité du corpus. En revanche, la méthodologie a été testée sur des échantillons, de manière à valider les principes retenus, de même que les propriétés et valeurs envisagées pour les annotations des chaînes. Ce document se contente donc de mentionner brièvement les résultats de ces tests effectués sur des échantillons.

A noter : il est possible de générer toutes les chaînes, ou seulement les chaînes d'au moins trois maillons. C'est l'un des paramètres de la macro de construction automatique des chaînes.

D.5.2 Annotation des proprietes des referents

Un schéma de type « chaîne » regroupe toutes les mentions référant à une même valeur de REF. On retrouve donc REF parmi les traits du schéma « chaîne », ce trait étant bien entendu rempli automatiquement. De même, un trait « nombre de maillons » est rempli automatiquement avec le nombre de mentions inclus dans chacune des chaînes.

Les autres traits sont à annoter manuellement. Ils concernent les propriétés des référents, à savoir le sexe (masculin pour un individu ou un groupe d'hommes, féminin pour un individu ou un groupe de femmes, indéterminé dans tous les autres cas) ; le nombre (singulier pour un individu ou un objet unique, groupe strict pour un groupe d'humains ou d'objets dont on connaît le cardinal exact, groupe flou dans les autres cas) ; ainsi que le type de référent. Pour ce dernier trait, nous retenons et adaptons les classifications utilisées usuellement pour les tâches de détection des entités nommées en traitement automatique des langues. L'important est d'annoter en tant que tels les humains, les objets concrets, les objets abstraits, les organisations, les lieux, les événements et ainsi de suite, comme le montre la copie d'écran de la figure 11, issue de l'un des tests dont il est question plus haut.

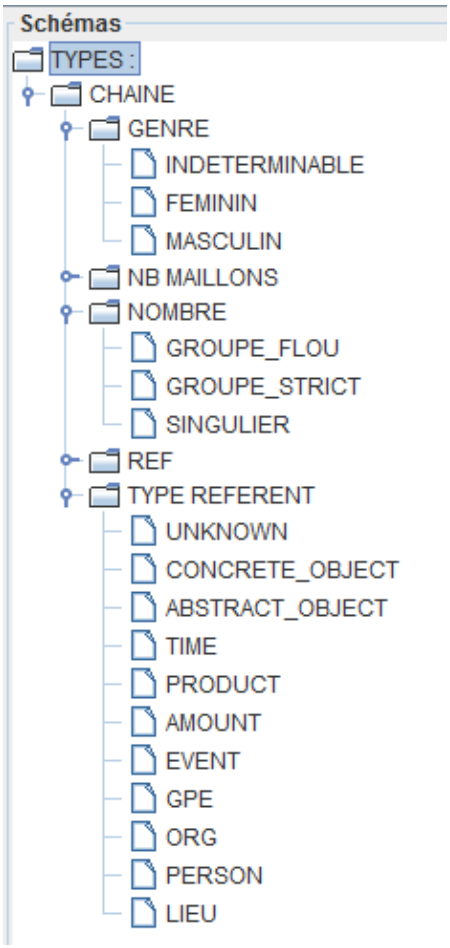

Figure 11 : structure d'annotation du type de schéma « chaîne ».

Nous terminons ainsi la partie qui concerne le manuel d'annotation – et par extension toute la procédure d'annotation du projet Democrat. Comme le nombre de pages dédiées à ce manuel n'est pas négligeable, il va de soi que les membres du projet ont dû être formés. De même, le consortium Democrat, et en particulier l'équipe en charge des développements de l'outil TXM, ont commencé un travail de réflexion et d'organisation de formations sur les nouvelles fonctionnalités d'annotation de TXM, ce qui fait l'objet de la section suivante.

E RECAPITULATIF DES FORMATIONS

E.1 FORMATIONS INTERNES AU PROJET

Compte tenu du nombre important d'informations contenues dans le manuel d'annotation, de nombreuses formations ont été réalisées en interne au projet. Elles concernent principalement deux aspects : 1. les subtilités de l'annotation des expressions référentielles (notamment leur repérage et leur délimitation) ; 2. l'utilisation des outils pour ce faire.

Concernant le premier point, plusieurs événements ont été programmés dans le projet Democrat : il s'agit tout d'abord des journées « MAD », nommées ainsi pour « Marathon Annotation Democrat ». Lors de ces journées, le principe était de mettre face à un texte à annoter le plus grand nombre de membres de Democrat, de manière à tester la procédure et à recueillir les questions et les difficultés rencontrées. De fait, trois journées ont eu lieu, une par site (Paris, Lyon, Strasbourg). Dans l'ensemble, la matinée a consisté à prendre connaissance du manuel et à annoter, et l'après-midi à discuter autour des difficultés rencontrées et des corrections à apporter au manuel d'annotation. Les journées MAD ont ainsi été les premières occasions pour l'ensemble des membres de Democrat de se retrouver en situation d'annotateur. Lors de chaque journée, une personne, voire deux, prenaient le rôle de formateurs.

Concernant le deuxième point, il s'est avéré que, depuis mars 2016, trois outils ont été utilisés pour annoter, les trois étant capables de produire le même type de fichier – on peut donc les considérer comme compatibles. Le premier est Analec, logiciel d'annotation de l'écrit développé depuis de nombreuses années au Lattice. Le second est SACR (mentionné dans le manuel d'annotation), script d'annotation dédié aux chaînes de référence et donc optimisé pour cette tâche. Il a été développé à Strasbourg par Bruno Oberlé et ne sert que pour Democrat, son caractère « optimisé » permettant une plus grande rapidité qu'Analec (dont la priorité est d'être universel, c'est-à-dire adapté à tous types d'annotations). Enfin, le troisième, venu plus tardivement dans le projet, est TXM. C'est d'ailleurs l'un des livrables du projet Democrat : l'extension de TXM avec des fonctions d'annotation inspirées d'Analec – en incluant le caractère universel de celui-ci – fait partie des objectifs (et des livrables à venir) de Democrat. TXM est utilisé par certains annotateurs depuis fin 2017 et surtout début 2018, sachant que la version utilisée n'est pas encore totalement terminée ni même stabilisée (le livrable TXM est à rendre en mars 2019). Il commence néanmoins à être utilisable, surtout pour la première phase d'annotation, à savoir celle des mentions.

Nous sommes ainsi en présence de trois outils. Chaque annotateur est libre d'utiliser celui qu'il souhaite. Dans chacun des laboratoires participant au projet, ont été nommés un référent, spécialiste d'un ou de plusieurs parmi ces trois outils. Frédérique Mélanie-Becquet s'est ainsi chargée de jouer ce rôle de formatrice-outil au Lattice, les outils qu'elle connaît et pour lesquels elle peut donner des formations étant Analec et TXM. Matthieu Quignard s'est lui aussi chargé de jouer ce rôle de formateur à Lyon (ICAR et IHRIM). TXM étant développé à Lyon, il est logique que les formations qui y ont été données concernent TXM. Enfin, Bruno Oberlé s'est chargé du rôle de formateur à Strasbourg, avec une focalisation sur SACR et TXM. Selon le laboratoire d'appartenance d'un membre du projet, les trois choix restent possibles, mais un suivi et des formations ne peuvent être assurés en direct que pour deux parmi ces trois choix. Cela n'a posé jusqu'ici aucun problème.

Il est à noter que toutes ces formations ont pris un temps conséquent. Non seulement d'un point de vue technique – problèmes d'installation, d'ergonomie, de bonne utilisation des fonctions proposées par l'outil, etc. – mais aussi d'un point de vue théorique, relevant de la nature complexe des annotations elles-mêmes. Annoter la référence et la coréférence s'est

avéré une tâche très complexe, qui a suscité de nombreuses questions et discussions. Des formations ont donc été données aux annotateurs tout au long de leur travail. Ces formations ont pris plusieurs formes (qu'il est difficile de quantifier en nombre d'heures) :

- mise en place et maintien d'un wiki des annotateurs, regroupant des exemples bizarres ou remarquables, des problèmes d'utilisation, ou encore des trucs et astuces pour l'annotation de certains phénomènes ;
- mise à jour en fonction des retours du manuel d'annotation, pour qu'il inclue des précisions techniques (copies d'écran). Compte tenu de l'existence des trois outils, il a fallu choisir un mode de fonctionnement qui ne conduise pas à trois manuels d'annotation différents, mais à un seul manuel avec des copies d'écran montrant des phénomènes quasiment « communs » aux trois outils, en tout cas permettant à l'annotateur de s'y retrouver quel que soit l'outil utilisé ;
- organisation de réunions et de séances de discussion entre annotateurs pour évoquer les cas particuliers rencontrés et échanger autour des critères de décision à appliquer dans ces cas.

Le dernier de ces trois points a sans doute été le plus chronophage, en particulier au début de l'annotation.

E.2 FORMATIONS POUR LA COMMUNAUTE

Comment aller au-delà du groupe formé par les membres de Democrat et proposer des formations à des étudiants voire des chercheurs de la communauté de la linguistique de corpus outillée et intéressés par la référence et la coréférence ?

Pour répondre à cette question, nous avons choisi deux voies :

1. La formation d'étudiants à la méthodologie de l'annotation manuelle de corpus, en prenant comme illustration l'annotation de la référence et des chaînes de référence. Pour ne pas mélanger trop d'informations et rendre la formation indigeste, il a été décidé de laisser de côté les aspects « outils », c'est-à-dire à la fois le choix de l'outil adéquat et les aspects ergonomiques des différents outils disponibles (en dehors même de TXM, qui est largement répandu dans la communauté). En 2016 et 2017, cette formation a pris la forme de 2 à 6 heures de cours dans un module dénommé « e-philologie » d'une formation organisée par Thierry Poibeau (directeur du Lattice et membre de Democrat) dans le cadre de PSL University (Paris-Science-et-Lettres). Les cours ont été donnés par Frédéric Landragin et Frédérique Mélanie-Becquet devant des étudiants de l'ENS Paris, l'ENC (Ecole Nationale des Chartes), l'EPHE (École Pratique des Hautes Études) et l'EHESS (Ecole des Hautes Etudes en Sciences Sociales). En 2018, le module est devenu partie prenante d'un master, le master « Humanités Numériques » de PSL. Democrat contribue donc à ce master orienté sur

les Humanités Numériques en y apportant des repères méthodologiques illustrés par des cas concrets rencontrés dans le projet (cours de 3 heures en 2018).

2. La formation d'étudiants et de chercheurs à la plateforme TXM, en orientant l'une des formations sur les possibilités d'annotation. Ce projet a été possible grâce au TGIR Huma-Num, plus précisément au consortium CORLI (« CORpus, Langues, Interactions ») dont font partie plusieurs membres de Democrat, notamment Serge Heiden à Lyon, Frédéric Landragin et Clément Plancq à Paris. L'événement a eu lieu lors de journées scientifiques « Explorer un corpus annoté », les 25 et 26 octobre 2017. Frédéric Landragin a présenté le projet Democrat et donné une introduction à l'annotation et sa méthodologie le 25 octobre, sous le titre « Le projet Democrat : annoter et explorer des chaînes de référence avec TXM » ; Serge Heiden a présenté TXM et organisé un atelier autour des fonctionnalités d'annotation issues de Democrat le 26 octobre, sous le titre « Annoter avec TXM ».

Ces deux initiatives ayant été bien perçues, il y a de grandes chances que nous les reconduisions pendant les années à venir.