Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Intégration des relations inter-référents dans l'annotation de la coréférence : modèle et application

Antoine Boiteau Yann Mathet Antoine Widlöcher
Normandie Université, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, FRANCE
<prenom.nom>@unicaen.fr

RÉSUMÉ

La disponibilité de corpus annotés en coréférence demeure une nécessité pour de nombreux travaux en linguistique et en TAL. Toutefois, si de tels corpus sont bien disponibles, une part importante repose sur des modèles d'annotation ne permettant d'encoder qu'une partie des informations liées aux phénomènes coréférentiels. Après avoir redéfini un modèle élargi de la coréférence, nous montrerons les bénéfices d'une annotation menée à deux niveaux, celui de l'inscription des occurrences dans le texte (le repérage des maillons des chaînes de coréférence, niveau largement exploré) et celui des structures du modèle référentiel inféré (la clarification des rapports entre les entités désignées, domaine largement passé sous silence). Nous présenterons ensuite l'environnement OPERA destiné à l'annotation selon ce modèle repensé, et une campagne menée pour le tester.

ABSTRACT

Integrating relation between entities when annotating coreference : model and application

Availability of annotated corpora on coreference is a requirement for many tasks in linguitics and NLP. However, even if such corpora are available, they mostly rely on annotation models which only consider some parts of coreferential phenomena. After redefining a wider model for coreference, we show benefits of using annotation at two levels, one concerning occurrences markup in the text (spotting the units of coreference chains, a widely discussed level) and the second concerning infered structures from the referential model (the characterization of connections between recognized entities, a largely unexplored level). We also introduce the annotation tool OPERA, which is designed for annotation according to this new model, and experiments for testing it.

MOTS-CLÉS : Modélisation de la coréférence, campagne d'annotation, outil d'annotation.

KEYWORDS: Coreference model, annotation campaign, annotation tool.

1 Introduction

L'importance des phénomènes coréférentiels, tant dans une perspective linguistique qu'en traitement automatique des langues, est bien connue (Schnedecker & Landragin, 2014) et de nombreuses applications en rendent l'analyse nécessaire comme condition *sine qua non* d'accès au sens des textes. Aussi, de nombreux travaux y ont été consacrés, dont des états de l'art récents (Sukthanker *et al.*, 2020; Labat & Aufrant, 2024) et des travaux nombreux, y compris dans notre communauté francophone (Lopez, 2023; Barletta, 2024), rendent bien compte, tant sur le terrain de leur étude linguistique que sur le terrain de la résolution automatique, à travers des approches très variées.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 437–450.

Quoi qu'il en soit de la diversité des approches, la disponibilité de corpus annotés de référence demeure souvent une nécessité, que ce soit pour une étude des phénomènes en contexte, pour l'entraînement ou le *fine tuning* de systèmes d'apprentissage, ou pour l'identification d'exemples d'intérêt pour le paramétrage de systèmes génératifs. Notre communauté a donc naturellement pris en charge l'établissement de tels corpus de référence, à l'occasion de campagnes dont les travaux de Schang *et al.* (2011), Lefeuvre *et al.* (2014) et Landragin (2016) constituent des exemples largement commentés et utilisés pour la langue française. Toutefois, il apparaît que la majorité des corpus disponibles repose sur des modèles d'annotation relativement simples, ne permettant d'encoder qu'une partie des informations liées à la coréférence. Cela résulte sans doute largement de contraintes volumétriques, qui obligent les responsables des campagnes à définir des instructions d'annotation assez simples, mais aussi, pour autant que nous puissions en juger, des spécificités des environnements d'annotation utilisés, dont l'expressivité détermine souvent les choix faits par ces responsables.

Après avoir redéfini en un sens assez large le périmètre possible des études coréférentielles, nous voudrions défendre ici l'idée que l'annotation de la coréférence pourrait tirer bénéfice d'une annotation à deux niveaux, celui de l'inscription des occurrences dans le texte (le repérage des maillons des chaînes de coréférence, niveau largement exploré) et celui des structures du modèle référentiel inféré (la clarification des rapports entre les entités auxquelles le texte fait référence, domaine largement passé sous silence). Nous présenterons ensuite l'environnement d'annotation OPERA qui est destiné à l'annotation selon ce modèle repensé de la coréference et commenterons pour finir les résultats d'une campagne d'annotation menée dans le but de tester cet environnement OPERA.

2 État de l'art

2.1 Chaînes de (co)référence et Référents

L'étude des phénomènes référentiels au sein des textes est un sujet de recherche prolifique depuis plusieurs décennies, en linguistique comme en TAL. On s'intéresse tout particulièrement ici à la représentation numérique de ces phénomènes et à la nature des liens qui les font naître.

L'élément primordial du phénomène référentiel est l'*expression référentielle*, un mot ou un groupe de mots établissant une connexion symbolique avec une entité particulière (Chastain, 1975), une unité textuelle qui *fait référence* à une entité, réelle ou fictionnelle, qui appartient au cadre du discours. Les expressions référentielles sont aussi souvent appelées *maillons*. En filant la métaphore, Chastain (1975) rassemble les maillons faisant référence à la même entité en une *chaîne*. Au sein de la communauté TAL on parlera, de manière interchangeable, de chaîne de référence, de chaîne de coréférence ou même simplement de coréférence (Landragin, 2021), même si ces notions ne sont pas linguistiquement équivalentes (Schnedecker, 2019). Nous faisons ici usage de la notion de *chaîne de référence* pour qualifier les ensembles de maillons et nommons *référent* l'entité désignée par un certain nombre de maillons, entité support de regroupement de ces derniers en une chaîne.

2.2 Modélisation de la coréférence dans la littérature

Dans la littérature, nous distinguons deux grandes familles de représentation des chaînes de référence. La figure 1 donne une intuition schématique des différentes approches commentées ci-après.

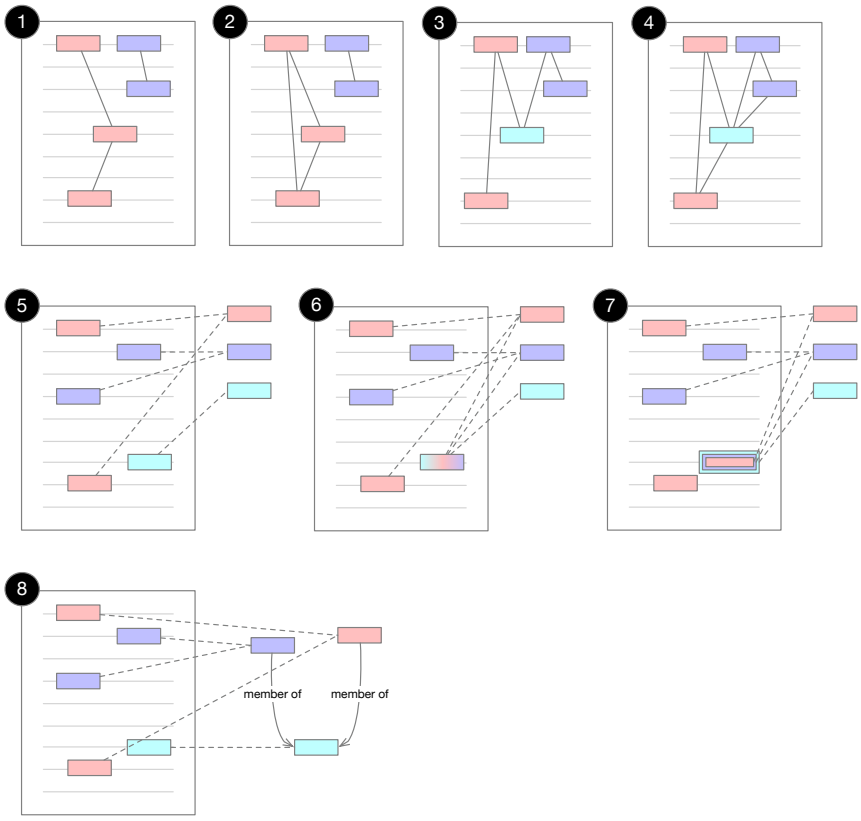

FIGURE 1 – Exemples de modélisations relationnelles (❶ à ❹), ensemblistes (❺ à ❼) et de notre proposition (❽). En bleu clair, des références à des groupes d'entités.

Le premier type de modélisation, que nous nommerons *modélisation relationnelle* (❶ à ❹), est au plus près de la notion de chaîne au sens où, lorsqu'un premier maillon m exprime une référence à une entité e et qu'un second maillon m' réfère à la même entité e, alors une relation symbolisant le phénomène de coréférence est annotée entre m et m'. La chaîne de référence émerge par l'assemblage des relations unissant un à un ses maillons. Dans cette optique on comprend la motivation de Schnedecker & Landragin (2014), qui proposent de ne qualifier la présence d'une chaîne de référence qu'à partir de 3 maillons, la seule relation de coréférence suffisant à caractériser le phénomène entre deux maillons, et le *singleton* ne pouvant compter ni comme chaîne, ni comme relation. Cette formalisation de la chaîne de référence par les relations inter-maillons se retrouve dans des corpus bien connus au niveau international, par exemple OntoNotes (Weischedel *et al.*, 2011) et ARRAU (Poesio *et al.*, 2024), ainsi que dans des corpus francophones tels que CO2 (Schang *et al.*, 2011) et ANCOR (Lefeuvre *et al.*, 2014), bien que ce dernier propose aussi une traduction de l'annotation des chaînes dans la seconde approche que nous abordons maintenant.

Le second type de modélisation identifié dans la littérature, que nous nommerons *modélisation ensembliste* (❺ à ❼), consiste à annoter uniquement le lien entre le maillon et son référent. La chaîne de référence est encodée par l'ensemble des maillons faisant référence à un même référent. Dans le processus d'annotation, cela se traduit généralement par l'ajout, à la caractérisation de chaque maillon d'une même chaîne, d'une même étiquette portant l'identifiant du référent associé. La particularité de cette approche est qu'elle dispense l'annotateur d'une quelconque mise en relation, et réduit son effort à une tâche de catégorisation. Democrat (Landragin, 2016) et Scolinter (Wolfarth *et al.*, 2017) sont des exemples de corpus annotés selon une approche de ce type.

2.3 Expressivité et limites de ces modélisations

Notons que ces deux familles de modélisation ne sont pas complétement imperméables l'une à l'autre. La modélisation relationnelle peut en effet être traduite en modélisation ensembliste en rattachant chaque maillon à son référent plutôt qu'à un autre maillon coréférent. Cependant, pour les campagnes ne rattachant pas naïvement chaque maillon au dernier maillon coréférent dans la lecture du texte mais plutôt vers une *tête* spécifique, la traduction implique une perte d'information. Le retour de la modélisation ensembliste à la modélisation relationnelle n'est alors pas garanti.

Bien que souvent employés dans les campagnes d'annotation de la coréférence, ces deux types de modèles ne sont pas exempts d'écueils. L'un des plus frappants à la lecture des manuels d'annotation de différentes campagnes est la problèmatique de l'annotation de référents particuliers que nous nommerons *groupes*, et qui expriment la réunion de plusieurs autres référents présents dans le discours (par exemple dans le cas d'un couple, d'une classe, etc.). Utilisant la modélisation relationnelle, le corpus ARRAU préconise d'annoter pour un maillon faisant référence à un groupe composé d'un certain nombre de référents de type individuel, une relation de coréférence vers un autre maillon jugé antécédent adéquat respectivement pour chaque référent inclus dans le référent de type groupe (❸). Il en résulte deux inconvénients majeurs : tout d'abord, il faudra répéter le processus à chaque maillon faisant référence à un référent de type groupe ; par ailleurs, les maillons qui feraient référence au même groupe ne présentent pas de relation de coréférence annotée entre eux, si bien que la chaîne de référence pour ce référent de type groupe n'est pas exprimée par l'annotation. Concernant l'approche ensembliste, le manuel d'annotation du corpus Democrat propose de déclarer, dans le cas où deux référents de type individuel A et B seraient inclus dans un troisième référent de type groupe C, de déclarer autant de chaînes distinctes via la structure de traits des maillons (❻). Ce faisant, il est ainsi possible de regrouper tous les maillons faisant référence à C en une chaîne, mais aucun lien n'est exprimé dans l'annotation entre les référents A, B et C ni entre leurs maillons respectifs, bien que le dernier soit *composé* des deux autres et que chaque maillon de C porte aussi la trace sémantique de A et B. Pour ce cas précis le corpus Scolinter propose une autre méthode d'annotation qui consiste à annoter autant de maillons sur le passage de texte exprimant la référence à un groupe qu'il y a de référents de type individuel inclus dans ce groupe (❼). Chaque maillon issu du même passage est donc lié respectivement à la chaîne d'un référent de type individuel différent. La solution proposée par le corpus Scolinter se rapproche de celle proposée par Zhou & Choi (2018) et partage aussi la même difficulté que l'approche relationnelle illustrée par le corpus ARRAU : la chaîne de référence du référent de type groupe n'apparaît plus en tant que telle dans l'annotation.

Par ailleurs, il est manifeste qu'une corrélation entre le modèle de la référence et l'outil d'annotation utilisé existe. Il est cependant difficile de savoir si ce sont les modèles d'annotation qui inspirent les outils d'annotation, ou si ce sont les outils disponibles qui contraignent les modèles d'annotation des campagnes. De nombreux outils d'annotation, génériques ou spécifiques à cette tâche, ont été employés pour établir les corpus de référence. Citons quelques outils bien connus de la communauté, dont les fonctionnalités, parfois partagées, ont pu nous inspirer : MMAX2 (Müller & Strube, 2006), ANALEC (Landragin *et al.*, 2012), GLOZZ (Widlöcher & Mathet, 2012), ou encore INCEpTION (Klie *et al.*, 2018). Nous observons dans quelques outils d'annotation plus récents, moins répandus et spécifiques à l'annotation de la coréférence, une tendance à proposer une liste des référents intégrée à l'interface et déportée à côté du texte, comme c'est le cas par exemple pour l'outil *ad hoc* Coref Annotator présenté par Aralikatte & Søgaard (2020), ou encore pour SACR [1], qui vise à faciliter la tâche d'annotation par l'utilisation, entre autres, de jeux de couleurs permettant d'identifier rapidement

1. https://boberle.com/projects/coreference-annotation-with-sacr/

les référents et leurs maillons liés (Oberle, 2018).

Mélanie-Becquet & Landragin (2014) listent les besoins que doivent satisfaire les outils d'annotation de la coréférence. Entre autres, ces outils « [doivent] permettre une construction simple et efficace du monde des référents » et l'« annotation des relations entre chacun des référents en présence ». Or, à notre connaissance, aucun outil de l'état de l'art ne permet de satisfaire cette seconde contrainte.

3 Un nouveau modèle pour l'annotation de la coréférence

3.1 Pourquoi un nouveau modèle ?

La justification de l'adoption d'un nouveau modèle passe par la réaffirmation de l'horizon fonctionnel de notre examen des phénomènes coréférentiels. Pour accéder au sens d'un texte, il est nécessaire de savoir à qui ou à quoi tel ou tel énoncé se rapporte. Si c'est le cas, la tâche consistant simplement à mettre en lien une occurrence avec l'entité à laquelle elle réfère immédiatement, intransitivement, ne sera pas suffisante. L'entité immédiatement désignée peut en effet entretenir avec d'autres référents des relations telles qu'un énoncé portant sur l'un de ces autres référents pourrait nous informer aussi, transitivement, et en fonction du type des entités et de leur relation, sur l'entité considérée. Si un énoncé porte par exemple sur un groupe d'individus, certaines propositions valables pour le groupe vaudront aussi pour ses membres, la réciproque étant parfois vraie aussi. Pour tenir compte de ce que nous nommerons *transitivité*, il faut donc soit permettre un rattachement d'une occurrence à plusieurs entités, soit permettre la caractérisation des relations entre les entités du modèle référentiel.

Une approche que l'on peut qualifier d'*extensionnelle* pourrait consister à associer à toute mention référentielle d'un énoncé l'ensemble des référents qui lui sont associés (❻ et ❼). Si elle permet d'apporter une certaine réponse à la question de l'annotation des groupes, elle présente cependant des inconvénients du point de vue de l'expressivité et d'un point de vue pratique. Du point de vue de l'expressivité, une approche extensionnelle est incompatible avec des formulations non exhaustives, qui sont pourtant fréquentes. Dans « Pierre et Marie sont membres du même club d'échecs », les membres du club ne se limitent aux seuls Pierre et Marie. De même, dans « l'effectif de l'équipe a été multiplié par deux », il n'est pas précisé quels membres ont été ajoutés. Le fait de disposer d'un référent associé au groupe en tant que tel, sans que la composition de ce dernier ne soit forcément complètement établie, résout naturellement ce problème. Du point de vue pratique, il devient rapidement impossible pour l'annotateur humain de tisser les liens associés à chaque mention d'un groupe, tant le travail devient fastidieux et difficilement visualisable, comme le suggèrent ❻ et ❼. Il semble alors indispensable de prendre appui sur un modèle, que nous qualifions d'*intensionnel*, permettant de référer à des entités qui ne relèvent pas nécessairement du niveau de l'entité individuelle, mais pouvant représenter notamment des groupes d'entités.

3.2 Relations inter-référents

Les modélisations relationelles et ensemblistes rencontrent des difficultés d'expressivité qui, selon nous, pourraient être en partie résolues en accordant aux annotateurs l'opportunité d'exprimer des relations directement entre les référents du discours. Nous proposons, dans ce but, d'étendre le modèle ensembliste, qui rattache chaque maillon à un référent, en lui adjoignant trois règles :

— chaque maillon possède un couple de bornes uniques dans le texte, i.e. il est impossible d'avoir deux maillons positionnés exactement au même endroit dans le texte ;
— les référents introduits dans le discours par une expression référentielle, y compris les singletons, sont déclarés et annotés *à côté* du texte dans un espace que nous nommons *modèle référentiel* ;
— entre deux référents distincts du modèle référentiel, on peut annoter autant de relations que nécessaire. Chacune peut être unidirectionnelle ou bidirectionnelle, et son type précise la nature du lien qui les unit, information nécessaire à, ou issue de, la compréhension du discours.

La possibilité d'annoter ainsi les liens entre les référents du discours, liens usuellement laissés de côté par les autres modèles, augmente fortement l'expressivité de l'annotation des phénomènes coréférentiels. Il résout les limites des modélisations ensemblistes et relationelles évoquées ci-dessus concernant l'annotation de référents de type groupe. En effet, la prise en charge du cas où deux référents de type individuel A et B sont inclus dans un référent de type groupe C peut consister à annoter, dans le modèle référentiel, les trois référents A, B et C, et à créer deux relations respectivement entre A et C et entre B et C, relations dont le type doit formaliser le lien d'appartenance entre un référent individuel et un référent de type groupe. De cette manière, chaque maillon concerné dans le texte peut précisément être attribué à A, B ou C.

3.3 Outil d'annotation des phénomènes référentiels

Dans l'intention de tester la faisabilité et l'intérêt du modèle d'annotation proposé en 3.2, un outil d'annotation *ad hoc* a été programmé par nos soins. À l'heure où nous écrivons ces lignes nous appelons cet outil encore en phase expérimentale OPERA (*Online Platform for Experimentation on Relational Annotation*). Cet outil d'annotation spécifiquement développé dans le but d'annoter les phénomènes coréférentiels s'inspire d'autres outils du domaine. La présentation générale et le focus sur le rattachement rapide de chaque maillon à un référent s'inspirent des travaux exploratoires menés avec SACR. La création de relations entre les chaînes de références est influencée par l'implémentation originale du méta-modèle URS (Unité-Relation-Schéma) dans GLOZZ (Widlöcher & Mathet, 2012), et plus particulièrement de l'annotation d'éléments de type *relation* entre des éléments *schémas* (les chaînes de référence de notre modèle pouvant s'assimiler à ces *schémas* dans la nomenclature URS). Le choix de programmer *ex nihilo* un nouvel outil d'annotation s'est imposé à nous parce qu'à notre connaissance aucune autre plateforme ne prend en charge l'annotation, dans un modèle référentiel, des relations inter-référents, avec l'expressivité que nous souhaitons. Si l'expressivité de GLOZZ s'approche de notre besoin par le support d'URS, les annotations qu'il permet restent trop près du texte, un travail séparé sur le modèle référentiel n'étant pas permis. Par ailleurs, étant donné le succès des outils d'annotation disponibles en ligne et leur facilité de déploiement, l'architecture d'OPERA propose elle aussi une interface d'annotation et de gestion de campagne *via* le web. Côté *front-end* les technologies du web HTML, CSS et JavaScript sont utilisées. Côté *back-end* le framework pour Python 3 Cherrypy [2] et des fichiers de configuration au format JSON sont employés.

Les caractéristiques principales d'OPERA sont :
— l'accessibilité et l'utilisabilité de l'outil dans un environnement web ne nécessitant aucune installation de la part des utilisateurs ;
— l'hébergement possible en local ou sur un réseau interne du serveur de l'application ;
— le contrôle simple de l'annotation pour l'utilisateur avec les fonctionnalités permettant de

2. https://cherrypy.dev/

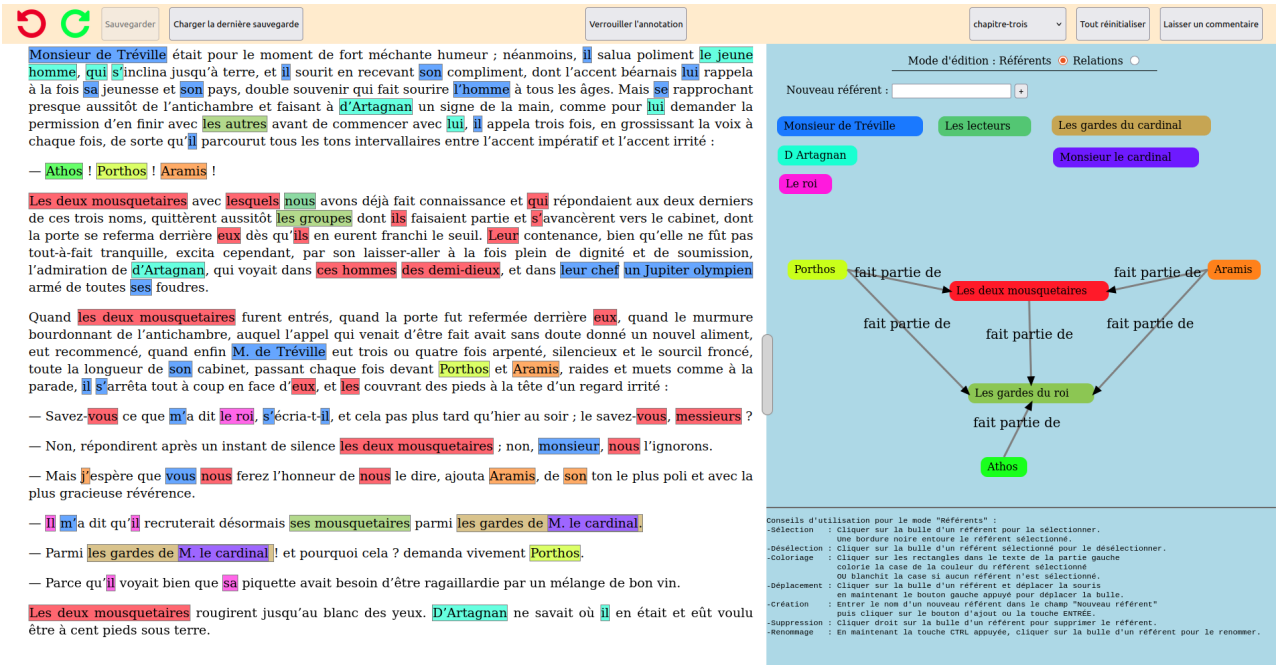

FIGURE 2 – Annotation de la coréférence en cours avec l'outil OPERA

défaire et refaire toutes les annotations, sauvegarder à tout moment, charger sa dernière sauvegarde, revenir à un document vierge, et enfin verrouiller son annotation pour soumission ;
— le support du balisage HTML des textes du corpus dans l'interface d'annotation pour représenter la stucture et la mise en forme du document ;
— la gestion et le verrouillage de fonctionnalités d'annotation au cas par cas à l'échelle d'un document ;
— la documentation contextuelle accessible directement dans l'interface de l'application.

L'outil OPERA dispose d'une interface ajustable divisée en deux parties illustrées par la figure 2. La partie gauche de l'interface présente le texte étudié et les maillons annotés. La partie droite illustre le modèle référentiel et présente deux modes, le premier pour la gestion des référents et le second pour la gestion des relations entre ces référents (voir la figure 3 en annexes).

4 Premières expérimentations avec OPERA

4.1 Annoter conjointement coréférence et liens entre référents

L'annotation de la coréférence dans les textes est une tâche non triviale souvent confiée à des annotateurs suffisament formés sur le sujet et il peut être difficile et coûteux de réunir de telles cohortes d'annotateurs. Pour cette raison et parce que nous voulions nous assurer que la compréhension de notre modèle était accessible à un large public, la cohorte composée pour nos premières expériences d'annotation avec OPERA a été consituée d'élèves francophones d'une classe de 4ème, impliquée dans ce travail pour mettre en évidence l'importance de la coréférence pour l'accès au sens des textes. Cette cohorte de 19 élèves a été subdivisée en deux sous-cohortes $C1$ et $C2$ de 8 et 11 annotateurs. Chaque élève travaillait individuellement après avoir visionné un guide vidéo expliquant les fonctionnalités de la plateforme.

La tâche d'annotation consistait à 1) relier les expressions référentielles *côté texte* à leur référent *côté modèle référentiel*, 2) établir entre les référents les relations nécessaires à la cohérence référentielle de leurs annotations. La cohorte $C1$ avait pour tâche additionnelle et préalable de 3) retrouver les personnages des textes et déclarer autant de référents correspondants. La cohorte $C2$ disposait dès le début de l'expérience de la liste complète des référents attendus déjà ajoutés au modèle référentiel. Deux textes de la littérature française devaient être annotés, la fable *Les Deux Mulets* de La Fontaine présentant peu de personnages (cinq référents dont un de type groupe), et un extrait du début du chapitre III des *Trois Mousquetaires* comprenant substantiellement plus de référents de type individuel et de type groupe (respectivement sept référents de type individuel et quatre de type groupe). Bien qu'OPERA permette de procéder à l'annotation des maillons dans le texte, le corpus était fourni avec tous ses marqueurs de référence pré-annotés manuellement, ici les syntagmes nominaux et pronoms faisant référence à un personnage du texte. La fonctionnalité de création de maillon était désactivée pour éviter les mauvaises manipulations et focaliser l'attention et le temps des élèves sur les autres tâches de l'expérience.

4.2 Résultats

Le principal objectif de l'expérimentation que nous avons menée était de mettre en lumière les capacités de 1) notre modèle à représenter de manière intelligible et fidèle les phénomènes coréférentiels et de 2) l'outil OPERA à permettre l'annotation selon un tel modèle. À la fin de l'exercice nous avons recueilli pour la cohorte 1 cinq versions annotées pour la fable (f pour la suite) et sept pour l'extrait de roman (r pour la suite). La cohorte 2 a produit 5 textes annotés pour f et 10 pour r. L'échantillon d'annotation recueilli est relativement petit mais suffisant pour identifier les tendances suivantes.

La cohorte 1 avait pour tâche particulière d'identifier lors d'une première lecture des textes les personnages présents et d'ajouter les référents relatifs à ces personnages au modèle référentiel en choisissant pour chacun d'entre eux un nom assez explicite. Pour les deux textes étudiés, tous les élèves ont trouvé et déclaré une partie des référents attendus. Plus précisément, pour f toutes les annotations sauf une présentent les cinq référents attendus. Pour r, plus long et possédant bien plus de référents, la majorité des élèves ont déclaré les référents les plus saillants (D'Artagnan, Porthos, Aramis et Monsieur de Tréville) mais aucun n'a trouvé la liste complète des référents attendus. Ces annotations démontrent que les annotateurs s'approprient correctement la création de référents et sont capables d'identifier les chaînes les plus saillantes d'un texte.

Après l'identification des référents en jeu, une tâche essentielle à l'identification des phénomènes coréférentiels est le rattachement des maillons à une chaîne. Avec OPERA, cela consiste simplement en une *coloration des maillons* : on repère le rattachement de chaque expression référentielle du texte à la couleur (ou l'absence de couleur) qui lui est attribuée et qui correspond à celle d'un référent dans la partie droite de l'interface. La tâche a globalement été comprise par l'ensemble des annotateurs mais présente des résultats variables en fonction des niveaux individuels de compréhension des textes. Des confusions multiples s'observent par exemple dans f, pour un énoncé tel que « Ami, lui dit son camarade » sans doute en raison des reprises croisées de deux référents masculins singuliers. Pour r, un cas intéressant est celui de la chaîne de référence *D'Artagnan* qui est introduite par le maillon « le jeune homme » semant l'ambiguïté sur l'identité réelle du référent, et qui n'est que plus tard explicitée par un premier maillon « d'Artagnan ». Certains annotateurs ont su identifier la continuité référentielle entre ces deux maillons, mais la majorité d'entre eux y ont vu deux chaînes différentes et ont procédé comme si un référent « jeune homme » existait et l'ont ajouté à la liste des

référents. Et ce même pour les annotateurs de la cohorte 2 ayant en leur possession la liste complète des référents attendus. La problèmatique du rattachement à la même chaîne pour deux maillons qui pourraient servir de tête sémantico-syntaxique se retrouve dans une moindre mesure dans f où les maillons « une troupe » et « l'ennemi » issus de la même chaîne ont pour une minorité d'annotateurs été annotés comme deux chaînes différentes. La différence de genre entre ces maillons a pu jouer un rôle déterminant. Cela révèle aussi, sans doute, une relative ambiguïté du passage.

Le dernier exercice consistait à annoter les relations entre référents. Les textes f et r ont été choisis parce qu'ils présentent au moins un référent de type groupe simple à identifier représentant l'union de deux autres référents de type individuel eux aussi remarquables. De plus, la catégorie de relation « fait partie de » était proposée aux annotateurs pour les inciter à lier les référents de type individuel au référent de type groupe. Un peu plus de la moitié des annotations recueillies ne présente pas de mise en relation entre les référents, ce qui peut être dû à plusieurs facteurs : la tâche était présentée en dernier, elle demandait plus de manipulations que les autres, et elle nécessitait une bonne compréhension du texte. Quelques annotateurs semblent cependant avoir bien compris la tâche, mais on observe une tendance de ces derniers à ajouter des types de relations non prévus par les consignes car sans lien direct avec les phénomènes coréférentiels (par exemple les relations « obéit à » ou « sert »).

Sur le plan technique, le retour d'expérience montre que les annotateurs les plus à l'aise sont tout à fait capables de se saisir des fonctionnalités d'OPERA pour repérer des référents dans un texte, de rattacher les maillons correspondant à ces référents et d'annoter des relations entre les référents. Le modèle semble donc tangible et l'outil exploitable, y compris auprès d'un large public. La qualité globale des annotations et les conclusions que nous pouvons en tirer ne sont cependant pas satisfaisantes pour l'intégralité des tâches, pour des raisons qui doivent maintenant être discutées.

4.3 Discussions

À l'issue de cette campagne, quelques points méritent d'être soulignés, qui nous guideront dans l'amélioration d'OPERA. Le premier concerne la nécessité de clarifier la tâche pour les annotateurs. On peut penser que cette clarification dépasse le cadre de l'environnement d'annotation et vaut surtout pour l'établissement des instructions d'annotation. Reste que, confronté au texte et au modèle référentiel, l'annotateur doit être guidé dans la compréhension de ces instructions lors du processus d'annotation lui-même. Nous avons en particulier remarqué que les annotateurs de la cohorte n'ayant pas à enrichir l'ensemble des entités référencées ont pourtant souvent cherché à ajouter des entités plutôt qu'à attacher des occurrences à des entités déjà clairement définies. Le cas de cataphore pourtant très simple à résoudre en début de texte (un « jeune homme » n'étant pas d'emblée identifié comme « d'Artagnan ») tend à montrer que les annotateurs ont pu penser que l'annotation devait refléter leur compréhension du texte au moment de leur première lecture de tel ou tel passage. Si la tâche consiste ainsi, par exemple, à annoter seulement après que l'ensemble du texte a été lu et compris, il conviendrait que l'interface impose cette démarche, pour éviter un processus d'annotation effectué au fil de la première lecture, par exemple en ne donnant accès aux fonctionnalités d'annotation que par un dispositif placé après le texte, pour signifier cet ordre des opérations.

Cette observation est intimement liée à la nécessité de pouvoir, pour une campagne donnée, brider les fonctionnalités inutiles et en conséquence potentiellement indésirables. Dans le cas qui nous occupe, que l'ajout de nouvelles entités ou de nouveaux types de relations entre entités soit techniquement possible, même si les instructions interdisaient explicitement d'en faire usage, n'a clairement pas suffi, et les annotateurs ont ajouté des éléments des plus hétéroclites. Selon un réflexe bien connu, toute

fonctionnalité offerte appelle à son usage et on peut même penser qu'un annotateur croira manquer à ses obligations en ne faisant pas systématiquement usage de toutes les possibilités offertes. Pour une annotation libre où l'utilisateur pourrait étendre l'ensemble des types de relations utilisables, cette difficulté se traduirait ainsi à coup presque certain par la prolifération des types les plus imprévisibles, les relations sémantiques les plus diverses risquant d'être regardées comme contribuant peu ou prou à la clarification référentielle et donc comme éligibles. Il convient donc selon nous de prédéfinir et de bloquer l'ensemble des types de relations utilisables.

5 Conclusion & perspectives

En conclusion de ce travail, nous souhaitons tout d'abord souligner que les décalages entre a) les modèles linguistiques de la coréférence, b) son intérêt applicatif et c) le périmètre retenu dans les campagnes d'annotation, méritent d'être discutés. En particulier, il nous semble utile de revenir à une définition assez couvrante du phénomène de la coréférence, en phase d'ailleurs avec son intérêt applicatif initial, soit, pour le dire brutalement : savoir à qui ou à quoi se rapporte tel ou tel énoncé. En réponse à la contrainte de transitivité qui résulte de cette exigence, et pour satisfaire au critère d'intensionnalité dont nous avons montré l'avantage, nous avons proposé un modèle à deux niveaux, celui du texte, comportant des occurrences d'expressions référentielles, et celui d'un modèle référentiel associé à l'interprétation du texte, et proposant de tisser des relations d'une part des expressions référentielles vers des instances du modèle référentiel, et d'autre part entre les instances du modèle référentiel elles-mêmes. Cette proposition nous semble être un enrichissement important par rapport aux approches effectives (c'est-à-dire disposant d'un outillage logiciel les mettant en œuvre) existantes à ce jour, qui permet non seulement un apport du point de vue de la pertinence linguistique, mais aussi de l'efficacité pratique des applications s'appuyant sur de telles annotations. Les échanges que nous avons pu avoir avec des responsables de campagnes d'annotation dans lesquelles cette richesse expressive n'est pas mise en œuvre nous font penser que c'est faute d'un outillage adapté que ces campagnes ne l'intègrent pas. En ce sens, nous espérons que la plateforme proposée dans cet article sera une contribution ou une inspiration pour les futures campagnes dans le domaine. Les premières expériences menées semblent montrer qu'elle propose un compromis intéressant entre expressivité et temps d'annotation. Reste que l'approche extensionnelle est plus présente dans la littérature et que, si l'approche intensionnelle que nous défendons est plus fidèle selon nous à la réalité décrite, il faudra néanmoins prolonger la discussion des rapports entre les expressivités respectives de ces deux approches.

Dans le prolongement de ce travail, une fois admise la nécessité d'annoter certaines relations entre entités du modèle référentiel, nous devrons continuer de travailler à l'épineuse question de la détermination de l'ensemble des relations pertinentes à retenir. Toute relation identifiable entre deux entités (sur un plan sémantique d'abord, mais aussi pourquoi pas en conséquence d'informations distributionnelles relatives à leurs occurrences en contexte) pourrait en droit être regardée comme potentiellement porteuse des effets transitifs évoqués ci-dessus, y compris dans les usages rhétoriques les plus vagues : si tel terme $T1$ a été lu au voisinage de tel autre terme $T2$, cette relation n'ouvrira-t-elle la possibilité qu'une proposition ultérieure se rapportant à $T2$ nous informe sur $T1$? Et ne tombera-t-elle pas alors sous le concept élargi de coréférence ? Au problème de cette délimitation, nous n'avons pas encore de réponse catégorique. Si les relations méréologiques semblent pouvoir faire consensus et pouvoir aussi nous occuper durablement, il est clair aussi que d'autres relations sont à la fois très fréquentes et indispensables à traiter pour la compréhension de textes même extrêmement simples et

aux effets rhétoriques négligeables. Nous pensons là, notamment, aux relations décrivant l'occupation par une entité référentielle d'une fonction liée à une autre entité : si X est président de Y, si X est représentant légal de Y, des énoncés portant sur X nous informeront souvent de façon décisive sur Y. Mais cette énumération ne nous convient guère et nous devrons trouver un critère permettant de caractériser plus strictement l'ensemble des relations à incorporer.

Le modèle élargi de la coréférence que nous avons en vue comporte aussi un élément essentiel actuellement non pris en charge dans l'environnement OPERA : la clarification de la portée textuelle et de l'ancrage temporel des relations établies entre les entités du modèle référentiel. Cela s'éclairera aisément avec un exemple emprunté au cas important des relations méréologiques. Au fil d'un texte, l'appartenance d'une entité à un groupe d'entités est susceptible de varier : tel membre rejoint le groupe, puis le quitte ; tel groupe se sépare de telle cohorte... Pour chaque relation établie dans le modèle référentiel, il faut donc pouvoir, si cette relation n'est pas stable au fil du texte, spécifier les bornes (parfois floues d'ailleurs) entre lesquelles la relation est supposée valoir. À la question de cette portée textuelle s'ajoute celle de la portée temporelle d'une relation : telle relation entre entités peut ne valoir que pour une période de temps limitée, entre des bornes définies par rapport à la temporalité des faits rapportés. Notons qu'un modèle d'annotation réduisant l'encodage d'une relation entre entités à l'encodage de multiples liens entre chaque maillon de la chaîne associée et les différentes entités désignées soit immédiatement soit transitivement (approche que nous avons qualifiée d'extensionnelle), permet la clarification de la portée textuelle (quoique de manière ergonomiquement très coûteuse) : tel maillon, à tel moment du texte, peut pointer vers les entités $E1$ et $E2$, quand tel autre maillon, à tel autre moment du texte, ne pointera par exemple que vers $E1$. En revanche, un tel modèle d'annotation ne permettra pas l'encodage de la portée temporelle. Le modèle que nous envisageons permettra au contraire de décrire de manière homogène la portée textuelle et la portée temporelle, par la seule caractérisation des liens entre les entités du modèle référentiel.

D'autres éléments identifiés au moins dans le cadre des études linguistiques devront aussi être pris en charge. Notre modèle prévoit ainsi en particulier le traitement de l'incertitude et de l'ambiguïté qu'étudient déjà des travaux tels que ceux de Delaborde (2020, 2021). Dans notre approche, l'incertitude et l'ambiguïté peuvent concerner soit le rattachement d'un maillon à une ou plusieurs entités du modèle référentiel, soit la liaison entre des entités, quand une incertitude ou une ambiguïté persiste. Un symbole particulier sera proposé pour matérialiser cette incertitude dans le modèle référentiel représenté par OPERA. Interposé entre les éléments reliés (entre le texte et le modèle référentiel ou au sein du modèle référentiel), il permettra soit le partage d'un branchement vers de multiples cibles, soit simplement l'expression d'une incertitude pour un branchement simple. Au-delà, s'imposerait aussi la prise en considération de l'expression des points de vue et de la modalité, sans lesquels les phénomènes coréférentiels sont souvent ininterprétables : telle entité ne serait qu'hypothétiquement partie de telle autre, telle personne ne serait membre de tel groupe que dans l'esprit de tel protagoniste... S'ouvre là un champ protéiforme d'une grande richesse sémantique, auquel les espaces mentaux de Fauconnier (1984) pourraient offrir un cadre théorique aussi élégant que difficile à intégrer.

L'implémentation des améliorations mentionnées présente des défis conséquents, notamment en ce qui concerne la lisibilité des annotations produites, pour lesquels nous devrons envisager des solutions. Une ouverture de l'outil OPERA à des tâches d'annotation portant sur d'autres types de relations est aussi envisageable et sera facilitée par l'influence du méta-modèle URS sur le développement.

Références

ARALIKATTE R. & SØGAARD A. (2020). Model-based Annotation of Coreference. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 74–79, Marseille, France : European Language Resources Association.

BARLETTA M. (2024). Annotation de la continuité référentielle dans un corpus scolaire – premiers résultats. In M. BALAGUER, N. BENDAHMAN, L.-M. HO-DAC, J. MAUCLAIR, J. G MORENO & J. PINQUIER, Éds., *Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 28–41, Toulouse, France : ATALA and AFPC.

CHASTAIN C. (1975). Reference and Context. *Minnesota studies in the philosophy of science*, **N° 7**, 194–269. Éditeur : University of Minnesota Press, Minneapolis.

DELABORDE M. (2020). *Analyse en corpus de chaînes de coréférence : la coréférence non-stricte à l'épreuve de la linguistique outillée*. Thèses, Université de la Sorbonne nouvelle - Paris III. Issue : 2020PA030073.

DELABORDE M. (2021). La coréférence floue dans les chaînes du corpus DEMOCRAT. *Langages*, **N° 224**(4), 47–65. Place : Paris Publisher : Armand Colin, DOI : 10.3917/lang.224.0047.

FAUCONNIER G. (1984). *Espaces mentaux : aspects de la construction du sens dans les langues naturelles*. Propositions. Paris : Édition de minuit.

KLIE J.-C., BUGERT M., BOULLOSA B., CASTILHO R. E. D. & GUREVYCH I. (2018). The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, p. 5–9 : Association for Computational Linguistics. Place : Santa Fe, USA.

LABAT L. & AUFRANT L. (2024). Évaluation de l'apport des chaînes de coréférences pour le liage d'entités. In M. BALAGUER, N. BENDAHMAN, L.-M. HO-DAC, J. MAUCLAIR, J. G MORENO & J. PINQUIER, Éds., *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, p. 397–409, Toulouse, France : ATALA and AFPC.

LANDRAGIN F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, (92), 11–15. Publisher : AFIA.

LANDRAGIN F. (2021). Le corpus DEMOCRAT et son exploitation. Présentation. *Langages*, **N° 224**(4), 11–24. Place : Paris Publisher : Armand Colin, DOI : 10.3917/lang.224.0011.

LANDRAGIN F., POIBEAU T. & VICTORRI B. (2012). ANALEC : a New Tool for the Dynamic Annotation of Textual Data. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 357–362, Istanbul, Turkey : European Language Resources Association (ELRA).

LEFEUVRE A., ANTOINE J.-Y. & SCHANG E. (2014). Le corpus ANCOR_Centre et son outil de requêtage : application à l'étude de l'accord en genre et nombre dans les coréférences et anaphores en français parlé. *SHS Web of Conferences*, **N° 8**, 2691–2706. DOI : 10.1051/shsconf/20140801359.

LOPEZ F. (2023). Approches neuronales pour la détection des chaînes de coréférences : un état de l'art. In M. CANDITO, T. GERALD & J. G. MORENO, Éds., *Actes de CORIA-TALN 2023. Actes*

des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), p. 101–113, Paris, France : ATALA.

MÉLANIE-BECQUET F. & LANDRAGIN F. (2014). Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques. *Langages*, **N° 195**(3), 117–137. DOI : 10.3917/lang.195.0117.

MÜLLER C. & STRUBE M. (2006). Multi-level annotation of linguistic data with MMAX2. In S. BRAUN, K. KOHN & J. MUKHERJEE, Éds., *Corpus Technology and Language Pedagogy : New Resources, New Tools, New Methods*, p. 197–214. Frankfurt a.M., Germany : Peter Lang.

OBERLE B. (2018). SACR : A Drag-and-Drop Based Tool for Coreference Annotation. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

POESIO M., CAMILLERI M., CARRETERO GARCIA P., YU J. & MÜLLER M.-C. (2024). The ARRAU 3.0 Corpus. In M. STRUBE, C. BRAUD, C. HARDMEIER, J. J. LI, S. LOAICIGA, A. ZELDES & C. LI, Éds., *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, p. 127–138, St. Julians, Malta : Association for Computational Linguistics.

SCHANG E., BOYER-PELLETIER A., MUZERELLE J., ANTOINE J.-Y., ESHKOL I. & MAUREL D. (2011). Coreference and anaphoric annotations for spontaneous speech corpos in French. In *DAARC'2011, 8th Discourse Anaphora and Anaphor Resolution Colloquium*, p. 9 pp., Faro, Portugal.

SCHNEDECKER C. (2019). De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence. *Les cahiers de praxématique*.

SCHNEDECKER C. & LANDRAGIN F. (2014). Les chaînes de référence : présentation. *Langages*, **N° 195**(3), 3–22. Place : Paris Publisher : Armand Colin, DOI : 10.3917/lang.195.0003.

SUKTHANKER R., PORIA S., CAMBRIA E. & THIRUNAVUKARASU R. (2020). Anaphora and coreference resolution : A review. *Information Fusion*, **N° 59**, 139–162. DOI : 10.1016/j.inffus.2020.01.010.

WEISCHEDEL R., HOVY E., MARCUS M., PALMER M., BELVIN R., PRADHAN S., RAMSHAW L. & XUE N. (2011). OntoNotes : A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation. Springer*, **N° 3**(3), 3–4.

WIDLÖCHER A. & MATHET Y. (2012). The Glozz platform : a corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on Document engineering*, p. 171–180, Paris France : ACM. DOI : 10.1145/2361354.2361394.

WOLFARTH C., PONTON C. & TOTEREAU C. (2017). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique. *Corpus*, (16). DOI : 10.4000/corpus.2796.

ZHOU E. & CHOI J. D. (2018). They Exist ! Introducing Plural Mentions to Coreference Resolution and Entity Linking. In E. M. BENDER, L. DERCZYNSKI & P. ISABELLE, Éds., *Proceedings of the 27th International Conference on Computational Linguistics*, p. 24–34, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

A Captures d'écran du logiciel expérimental OPERA

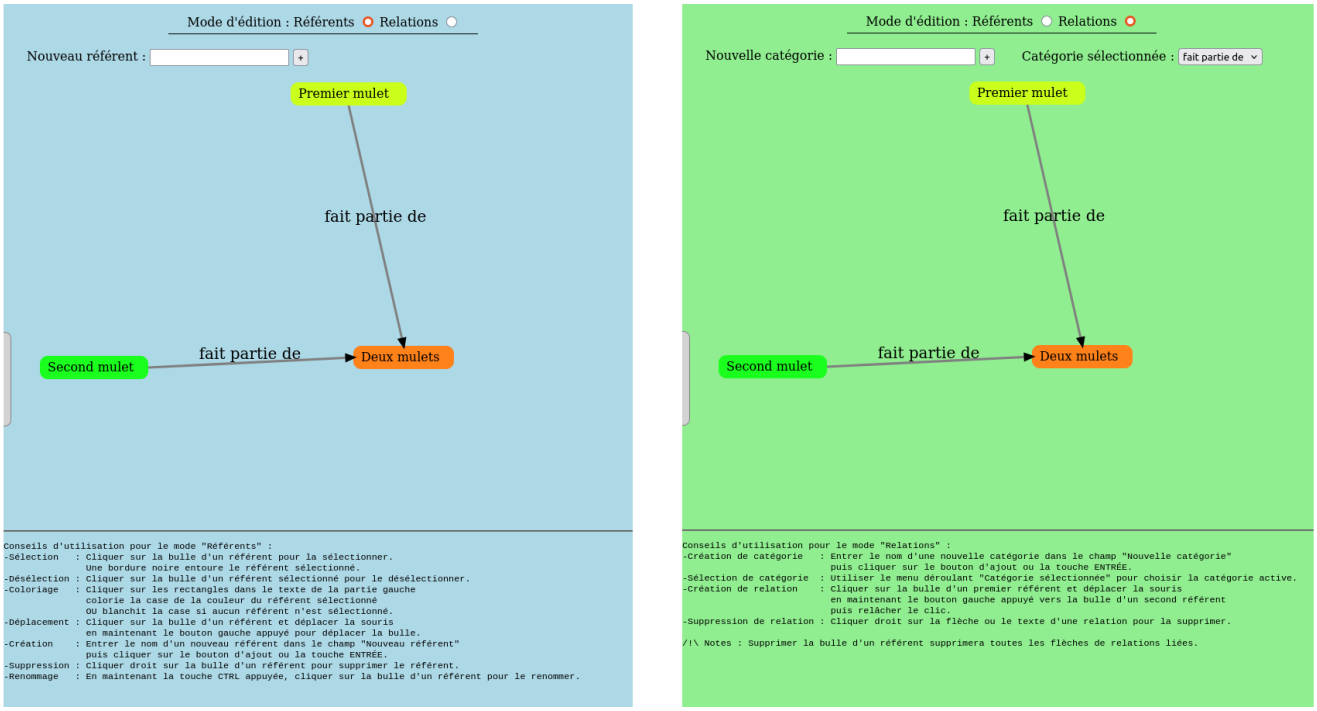

FIGURE 3 – Captures d'écran du modèle référentiel et de la documentation contextuelle de l'outil OPERA (en mode édition des référents à gauche et édition des relations à droite)